IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue1, November 2023

A Survey Paper on Moving Object Detection Using Deep Learning

[1]Aditi Maheshwari, [2]Anuj Badkat, [3]Milind Ankleshwar,[4]Reshma Sonar
Student, Department of Artificial Intelligence & Machine Learning[1,2]
Found Director, MASS IT Solutions LLP, Pune[3]
Associate Professor, Department of Artificial Intelligence & Machine Learning[4]
ISBM College of Engineering, Pune, Maharashtra, India[1,2,4]

Abstract: *Moving object detection in Python using deep learning is a powerful technique for accurately identifying and localizing moving objects in images or videos. By leveraging pre-trained models like YOLO or SSD, developers can implement this task efficiently. The Python implementation allows for customization and extension to handle real-time video streams and complex scenarios. This approach is valuable for researchers, practitioners, and enthusiasts interested in moving object detection using deep learning. Deep Convolution Neural Networks are leveraged to detect more precise coordinates and identify the category of objects. This survey paper provides study of various methodologies for object detection. This paper provides systematic analysis of various existing object detection techniques with precise and arranged representation.*

Keywords: Object Detection, Deep Learning, YOLO, SSD

I. INTRODUCTION

In general, the term Object recognition is used to describe a computer vision tasks that involve identifying objects in digital images. Image classification refers to process of predicting the class of an object. Object localization identifies the location of object in an image and draw bounding box around the object. Object detection, the focus area of this survey paper, often combines both these tasks of classification and localization in an image. Examples of object detection are face detection, pedestrian detection, etc.

Moving objects often convey the essential meaningful information in videos. Automatic detection of moving objects in videos is the first crucial stage for various potential applications, such as pedestrian and vehicle tracking, action and event recognition, annotation of video archives, etc. Several sophisticated works for detecting moving objects have been presented for stationary camera. The goal of moving object detection is to distinguish the foreground, which represents the moving objects of interest, from the static background. By detecting and tracking moving objects, it becomes possible to analyse their behaviour, classify them, count them, or take appropriate actions based on their presence or absence.

Moving object detection can be implemented using various technologies and techniques, including traditional computer vision algorithms, machine learning approaches, and deep learning models like Convolutional Neural Networks (CNNs). Depending on the complexity of the application and the desired accuracy, different methods and algorithms can be employed. Moving object detection has numerous practical applications, such as traffic monitoring, video surveillance, people tracking, action recognition, and more. It enables systems to understand and interpret the dynamic world captured in videos or image sequences, making it an essential component in many computer vision systems.

II. LITERATURE SURVEY

VM-MODNet: Vehicle Motion aware Moving Object Detection for Autonomous Driving.
Hazem Rashed,Ahmad El Sallab, LuqmanAli, Wasif Khanand Nakhon Poovarodom. [1]

ABSTRACT: Moving object Detection (MOD) is a critical task in autonomous driving as moving agents around the ego-vehicle need to be accurately detected for safe trajectory planning. It also enables appearance agnostic detection of objects based on motion cues. There are geometric challenges like motion parallax ambiguity which makes it a difficult

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue1, November 2023

problem. In this work, we aim to leverage the vehicle motion information and feed it into the model to have an adaptation mechanism based on ego-motion. The motivation is to enable the model to implicitly perform ego-motion compensation to improve performance. We convert the six degrees of freedom vehicle motion into a pixel-wise tensor which can be fed as input to the CNN model. The proposed model using Vehicle Motion Tensor (VMT) achieves an absolute improvement of 5.6% in mIoU over the baseline architecture. We also achieve state-ofthe-art results on the public KITTI MoSeg Extended dataset even compared to methods which make use of LiDAR and additional input frames. Our model is also lightweight and runs at 85 fps on a TitanX GPU.

MODETR: Moving Object Detection with Transformers
Eslam Mohamed, Ahmad El Sallab. [2]
ABSTRACT: Moving Object Detection (MOD) is a crucial task for the Autonomous Driving pipeline. MOD is usually handled via 2-stream convolutional architectures that incorporates both appearance and motion cues, without considering the interrelations between the spatial or motion features. In this paper, we tackle this problem through multi-head attention mechanisms, both across the spatial and motion streams. We propose MODETR; a Moving Object DEtectionTRansformer network, comprised of multi-stream transformer encoders for both spatial and motion modalities, and an object transformer decoder that produces the moving objects bounding boxes using set predictions. The whole architecture is trained end-to-end using bi-partite loss. Several methods of incorporating motion cues with the Transformer model are explored, including two-stream RGB and Optical Flow (OF) methods, and multi-stream architectures that take advantage of sequence information. To incorporate the temporal information, we propose a new Temporal Positional Encoding (TPE) approach to extend the Spatial Positional Encoding (SPE) in DETR. We explore two architectural choices for that, balancing between speed and time. To evaluate the our network, we perform the MOD task on the KITTI MOD [6] data set. Results show significant 5% mAP of the Transformer network for MOD over the state-of-the art methods. Moreover, the proposed TPE encoding provides 10% mAP improvement over the SPE baseline.

Moving Objects Detection with Freely Moving Camera via Background Motion Subtraction
Yuanyuan Wu, Xiaohai He, Truong Q. Nguyen. [3]
ABSTRACT: Detection of moving objects in a video captured by a freely moving camera is a challenging problem in computer vision. Most existing methods often assume that the background can be approximated by dominant single plane/multiple planes, or impose significant geometric constraints on background, or utilize complex background/foreground probabilistic model. Instead, we propose a computationally efficient algorithm which is able to detect moving objects accurately and robustly in a general 3D scene. This problem is formulated as a coarseto-fine thresholding scheme on the particle trajectories in the video sequence. First, coarse foreground region is extracted by performing reduced singular value decomposition (RSVD) on multiple matrices which are built from bundles of particle trajectories. Next, the background motion of pixels in coarse foreground region are reconstructed by a fast inpainting method. After subtracting the background motion, the fine foreground is segmented out by an adaptive thresholding method which is capable of solving multiple moving objects scenarios. Finally, the detected foreground is further refined by the mean-shift segmentation method. Extensive simulations and comparison to state-of-the-art methods verify the effectiveness of the proposed method.

Unsupervised Moving Object Detection via Contextual Information Separation
Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, Stefano Soatto. [4]
ABSTRACT: We propose an adversarial contextual model for detecting moving objects in images. A deep neural network is trained to predict the optical flow in a region using information from everywhere else but that region (context), while another network attempts to make such context as uninformative as possible. The result is a model where hypotheses naturally compete with no need for explicit regularization or hyper-parameter tuning. Although our method requires no supervision whatsoever, it outperforms several methods that are pre-trained on large annotated datasets. Our model can be thought of as a generalization of classical variational generative region-based segmentation,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue1, November 2023

but in a way that avoids explicit regularization or solution of partial differential equations at run-time. We publicly release all our code and trained networks.

Moving Object Detection for Event-based Vision using k-means Clustering
Anindya Mondal, Mayukhmali Das. [5]
ABSTRACT: Moving object detection is a crucial task in computer vision. Event-based cameras are bio-inspired cameras that mimic the working of the human eye. Unlike conventional frame-based cameras, these cameras pose multiple advantages, like reduced latency, HDR, reduced motion blur during high motion, low power consumption, etc. However, these advantages come at a high cost, as event-based cameras are sensitive to noise and have low resolution. Moreover, for the lack of useful visual features like texture and colour, moving object detection in these cameras becomes more challenging. Our proposed method uses k-Means clustering for detecting moving objects in event-based data. We further compare the proposed method against state-of-the-art algorithms and show performance improvement over them.

A Survey on Moving Object Detection and Tracking Based On Background Subtraction
Rahul Dutt Sharma, Subham Kumar Gupta. [6]
ABSTRACT: Detecting Moving object is a task to identify the motion of objects in a specific region/area. Over the last some years, moving object detection has received much attention due to its wide range of applications like human motion analysis, event detection, video surveillance, robot navigation, anomaly detection, traffic analysis and security, video conferencing etc. Video surveillance systems mostly deal with the tracking and classification of moving objects. The most common processing steps of motion detection for video surveillance includes detection of motion, modeling of environment, classification and object detection, behavior understanding and activity recognition. The main aim of this paper is to review recent developments and to analyze the open direction in visual surveillance systems for near future.

III. OBJECTIVE OF CNN

Convolutional Neural Networks (CNN, or ConvNet) are a type of multi-layer neural network that is meant to discern visual patterns from pixel images. In CNN, 'convolution' is referred to as the mathematical function. It's a type of linear operation in which you can multiply two functions to create a third function that expresses how one function's shape can be changed by the other. In simple terms, two images that are represented in the form of two matrices, are multiplied to provide an output that is used to extract information from the image.

IV. DEEP CONVOLUTIONAL NEURAL NETWORK IN MOVING OBJECT DETECTION

Deep Convolution Neural Networks (DCNNs) for object detection have gained a lot of interest for their powerful learning ability. By learning parameters themselves, it can achieve a high degree of accuracy. State of-the-art object detection networks comprise RCNN and its variants, SSD and its variants, YOLO and its variants, etc. RCNN and its variants are based on region proposal, which is accurate but time-consuming. YOLO and its variants [27]–[31] are known because of their fast speed and high efficiency. SSD and its variants blend the advantages of these two methods.

DCNN architectures used for moving object detection

YOLO

YOLO (You Only Look Once) is a popular object detection algorithm that uses deep convolutional neural networks (CNNs) to perform real-time object detection in images and videos. YOLO is specifically designed for object detection tasks. It can identify and locate multiple objects within an image or video frame by drawing bounding boxes around them. It is known for its speed and accuracy in detecting multiple objects simultaneously.
YOLO uses a deep convolutional neural network architecture as its backbone. The architecture is typically based on popular CNN models such as Darknet or ResNet. These models are pre-trained on large image datasets like ImageNet to learn powerful features. Unlike traditional object detection algorithms that use multi-stage pipelines, YOLO performs

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue1, November 2023

object detection in a single forward pass of the neural network. This makes it extremely fast, allowing real-time applications.

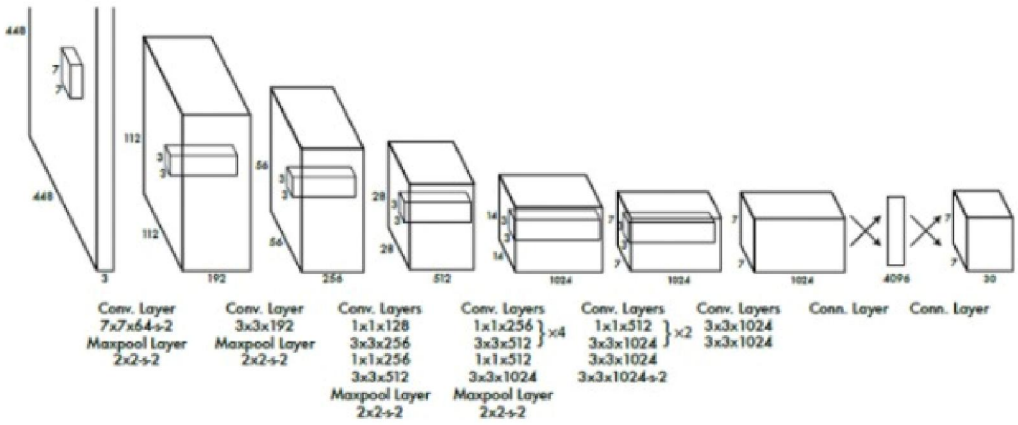
YOLO has several versions, such as YOLOv1, YOLOv2, YOLOv3, and YOLOv4, each with improvements in accuracy and speed.

Architecture of YOLO:-

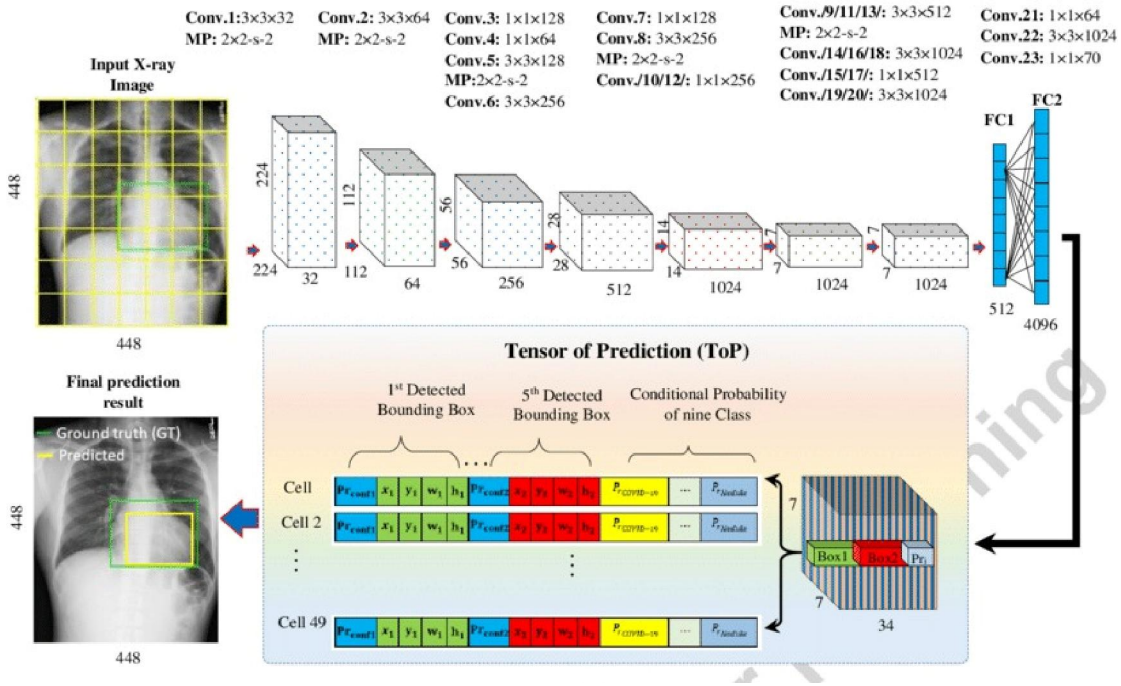
YOLO has overall 24 convolutional layers, four max-pooling layers, and two fully connected layers.

Input: Resizes the input image into 448x448 before going through the convolutional network.

- A 1x1 convolution is first applied to reduce the number of channels, which is then followed by a 3x3 convolution to generate a cuboidal output.
- **Feature Extraction**: The input image is passed through several convolutional layers in the backbone network to extract hierarchical features at different scales. These features capture both low-level and high-level information about the objects in the image.
- **Grid Division**: YOLO divides the input image into a grid of cells. The size of the grid can be adjusted based on the desired level of detection accuracy. Each grid cell is responsible for predicting bounding boxes and class probabilities for the objects present in that cell.
- **Anchor Boxes**: YOLO uses anchor boxes to handle objects of different sizes and aspect ratios. Anchor boxes are pre-defined bounding box shapes that are placed at each grid cell. The network predicts offsets and scales relative to these anchor boxes to accurately localize objects.
- **Predictions**: In each grid cell, YOLO predicts multiple bounding boxes along with corresponding class probabilities. Each bounding box consists of coordinates (x, y, width, height) relative to the anchor box, and each class probability represents the likelihood of an object belonging to a particular class.
- **Non-maximum Suppression**: After the predictions are generated, YOLO applies non-maximum suppression (NMS) to remove redundant detections. NMS selects the most confident bounding box among overlapping detections based on a predefined threshold. This helps eliminate duplicate detections and keeps only the most accurate bounding boxes.
- **Output**: The final output of YOLO is a set of bounding boxes, each associated with a class label and confidence score. These bounding boxes represent the detected objects in the input image.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue1, November 2023



R-CNN

RCNN stands for Region-based Convolutional Neural Network. It is the original version of RCNN and consists of two main steps. First, it generates region proposals using selective search or a similar algorithm. Then, each region proposal is passed through a CNN to extract features, which are then used for object classification and localization.

Fast R-CNN is an improvement over R-CNN in terms of speed. Instead of feeding each region proposal individually to the CNN, Fast R-CNN shares the convolutional features across all proposals. This reduces the computation time and allows for faster object detection.

Architecture of RCNN (Region-based Convolutional Neural Networks)

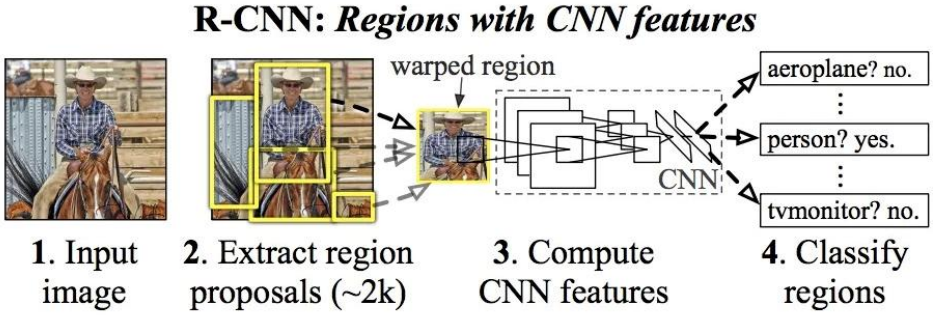
It consists of two main steps:

Region Proposal

Object Classification/Localization.

Region Proposal: RCNN uses a selective search algorithm to generate approximately 2000 region proposals from the input image. These region proposals are potential bounding boxes that may contain objects of interest.

Object Classification and Localization: Each region proposal is passed through a convolutional neural network (CNN) to extract features. The extracted features are then used for object classification and localization. The CNN is typically pre-trained on a large dataset, such as ImageNet, to learn generic visual representations.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue1, November 2023

SSD [Single Shot MultiBox Detector]

SSD is designed for object detection in real-time. Faster R-CNN uses a region proposal network to create boundary boxes and utilizes those boxes to classify objects. While it is considered the start-of-the-art in accuracy, the whole process runs at 7 frames per second. Far below what real-time processing needs. SSD speeds up the process by eliminating the need for the region proposal network.

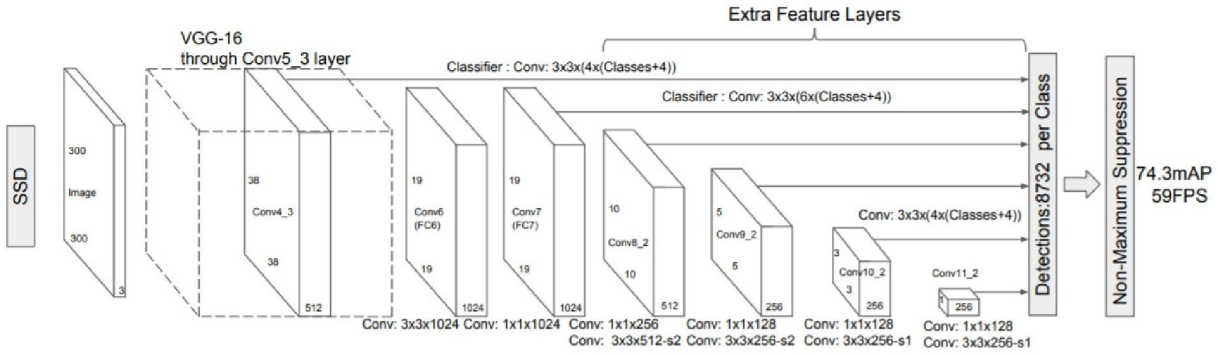
It is designed to be fast and efficient, making it suitable for real-time applications.

The SSD architecture consists of a backbone model, which is typically a pre-trained convolutional neural network (CNN) such as VGGNet or ResNet. The backbone model is used to extract features from the input image, which are then fed into a set of convolutional layers to predict the bounding boxes and class probabilities for different objects.

The SSD object detection composes of 2 parts:

Extract feature maps

Apply convolution filters to detect objects.



V. FRAMEWORK PROPOSED

Coarse-to Fine Grained Framework:

- Coarse-to-fine grained moving object detection framework combines moving detection with DCNNs.
- An efficient algorithm is proposed to detect the connected regions.
- The structure of the tiny version of YOLOV3 is modified to make the detection faster.

COARSE-GRAINED DETECTION:- In the coarse-grained detection step, filtering and mathematical morphology are also performed to reduce the adverse effect of noises. Firstly, image frames are filtered by lowpass filter to eliminate the high frequency noises. After that, moving detection algorithm is performed to detect motion. Finally, mathematical morphology (opening operation) is used to further suppress ill effects of noises.

In the high resolution scenes, we choose frame difference as the moving detection algorithm, which is simple to implement and is more responsive to almost all movements. Coarse-grained detection stage is suitable for high-resolution video sensing.

CONNECTED REGION DETECTION:- Connected region detection refers to the process of identifying and detecting regions or areas that are connected or contiguous in an image or data set. This type of detection is often used in computer vision and image processing tasks to analyze and understand the spatial relationships between different regions.

The algorithm considers different types of connectivity to define the relationship between pixels. The most common connectivity types are 4-connectivity and 8-connectivity. In 4-connectivity, pixels are considered connected if they share a common edge, while in 8-connectivity, pixels are considered connected if they share a common edge or corner.

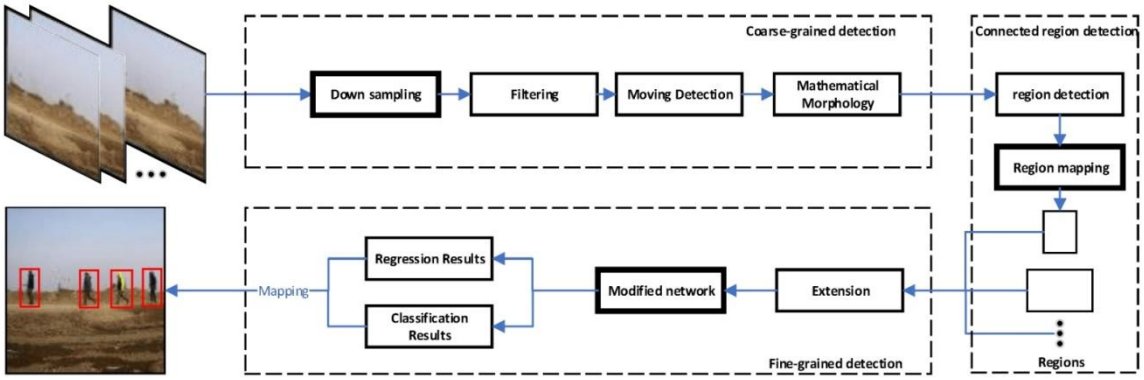
The algorithm then iterates through each pixel in the image and assigns labels to connected regions. It starts by initializing a label for the first foreground pixel encountered and then propagates the label to its connected neighbors. This process continues until all connected regions are labeled.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue1, November 2023

FINE-GRAINED DETECTION:- After coarse-grained detection and region extraction, it is much easier to detect objects. Fine- grained detection is closely related to fine-grained image classification, where the goal is to classify images into subcategories within a larger category. The focus is on recognizing and distinguishing between different instances of the same object category.

These models are capable of learning intricate features and patterns that are crucial for distinguishing between similar subcategories.

In the context of the fine-grained detection, there is an issue that moving objects obtained by coarse-grained detection is likely incomplete caused by noises. Therefore, before detected by the network, we extend the regions obtained in the coarse-grained detection to ensure the integrity of the objects. A method is proposed to extend the region according to the prior knowledge of the anchors in modified YOLOV3.Therefore, we use YOLOV3 for fine-gained detection to get faster speed.



VI. CONCLUSION

In conclusion, the survey provides a comprehensive review of deep learning-based object detection frameworks and approaches. The paper discusses the importance of using powerful backbone networks to extract rich features for object detection. The study leveraged the power of advanced machine learning techniques, particularly Convolutional Neural Networks (CNNs), to automate the process of detecting.

REFERENCES

[1]. VM-MODNet: Vehicle Motion aware Moving Object Detection for Autonomous Driving. Hazem Rashed,Ahmad El Sallab, Luqman Ali, Wasif Khanand Nakhon Poovarodom., July 2021.
[2]. MODETR: Moving Object Detection with Transformers, Eslam Mohamed, Ahmad A. Al Sallab, June 2021.
[3]. Moving Objects Detection with Freely Moving Camera via Background Motion SubtractionYuanyuan Wu, Xiaohai He, Truong Q. Nguye, IEEE. 2015.
[4]. Unsupervised Moving Object Detection via Contextual Information SeparationYanchao Yang, Antonio Loquercio, Davide Scaramuzza, Stefano Soatto, IEEE. April 2019.
[5]. Moving Object Detection for Event-based Vision using k-means Clustering. Anindya Mondal, Mayukhmali Das. 2020.
[6]. A Survey on Moving Object Detection and Tracking Based On Background Subtraction. Rahul Dutt Sharma, Subham Kumar Gupta, 2018.
[7]. A survey paper on object detection methods in image processing, manishavashisht; brijeshkumar, july 2020.