Deep Learning based Object Detection Techniques in Videos

[1]Kusuma S, [2]Kiran P and [3]Pavan Kumar S Kulkarni
Ramaiah Institute of Technology, Bengaluru, Karnataka 560054, India.
[1]kusumas.phd@gmail.com, [2]kiranpmys@gmail.com, [3]pavankumar.1rn16is065@gmail.com

ArticleInfo
R. Arulmurugan et al. (eds.), *First International Conference on Machines, Computing and Management Technologies*, Advances in Intelligent Systems and Technologies
Doi: https://doi.org/10.53759/aist/978-9914-9946-0-5_3

Abstract - Deep learning technology is often used for object detection. It has received attention recently because to the intimate connections between object detection, video analysis, and picture understanding. The goal of object detection has been pursued using a variety of models, and this is immensely beneficial to humanity. The most recent technical developments have helped the computational experiments, which would not have been conceivable if they had been tried using the conventional techniques. The powerful approaches employed in deep learning can show noticeably higher efficiency when compared to conventional designs and architectures. Numerous strategies and techniques have been used in deep learning to boost accuracy, and their drawbacks have also been somewhat addressed in order to lessen them. This study's main objective is to give an overview of several object detection procedures and approaches based on deep learning. Additionally, it lists the benefits and drawbacks of various object identification systems based on their potential applications and limitations.

Keywords – Deep Learning, Videos, Object Detection, RCNN, MRCNN.

I. INTRODUCTION

Researchers have made significant advancements in the field of object detection. The object's field There are numerous uses and approaches for detection. To find objects in digital photos, object detection is frequently utilised. Image classification is used to anticipate the object's class. Computer vision is an essential tool for achieving object classification, detection, and localisation. Models are trained on enormous datasets to achieve detection and categorise them to the present class of objects. Similar to the detection task is the localization task. Deep models don't require a specific regression or classifier tool. Object detection focus on the detection of a variable number of object classes, localisation concentrates on the identification of a fixed number of object classes. Applications for object detection include robotics, security systems, human-computer interaction, forensic operations, etc. Essentially, there are two phases involved in the detection of objects: training and testing.

To train the model, large datasets were used; as more images are added, the model is better able to recognise objects that belong to specific classes and better understands the class of items in general. The model's intended behaviour is verified during the testing phase. Feature extraction is a technique used by computers to teach them how to recognise features in objects. Every job that deep learning completes is optimised in an effort to improve the model's accuracy. The effectiveness and limitations of different object detecting systems are analysed here.

II. REVIEW OF LITERATURE

It helps to comprehend prior attempts to study object detection when a variety of research papers are evaluated for literature and object detection approaches are systematically analysed. The evaluation also emphasises the subject-matter specialists' exceptional contributions. Authors have made significant progress by addressing the problems and limitations associated with earlier developed object detection algorithms and seeking to work around them in subsequent investigations. To systematically arrange the strategies, a number of IEEE papers have been looked at as part of the survey.

Numerous methods have been looked at in [1], including Single Shot Multi box Detector (SSD), RCNN, MRCNN, over feat, SPP-net, YOLO, and FPN. Sliding windows, a simple approach, focuses on creating windows size, smaller than the actual size. When the full image has been looked at, the window's size is enlarged. This process is repeated using the new windows up until the condition to stop will be met.

The paper demonstrates how easy it is to apply the sliding window method. The fundamental objective of MRCNN is to extract as many features as possible from the diverse geographical proposals. Then, these characteristics are simply mixed. There are many areas that are taken into account for an object proposition, including half regions, core regions, boundary regions, contextual regions, and others.

The paper explains the ease of use of the sliding window method. The fundamental objective of MRCNN is to extract the maximum number of characteristics from the regional proposals that originate from everywhere. Then, these qualities are

just blended. In addition to the many different regions that are taken into account for an object proposition, there are also half regions, core regions, boundary regions, contextual regions, and others.

Short for Single Shot Multi box Detector, SSD. It feeds boxes with various scales to CNN's various levels, enabling each layer to forecast objects based on the scale value. The scale value is merely an image-furnishing parameter. Although SSD produces a sufficient amount of higher-level characteristics for larger objects, it is insufficient for smaller items. To increase the accuracy, a solid new generalizable strategy that can deliver rich semantics at all levels is required.

FPN, a further method being researched, is a feature extractor that can be used to modify the extractors of other detectors. As each layer is added to the system, the semantic values get richer. High resolution layers can be created using semantically rich layers. The semantic layers have a significant impact on FPN's performance, and one of its challenges is how it connects. This study also mentions the De-Net technique; whose main goal is to predict object candidate proposals by including a provision for bounding box corner estimate.

The key benefit is that it is substantially quicker than many other object detection systems and doesn't call for any specified anchors. The fact that this method takes more time to generate corners and evaluate the base network is one issue with it. The paper makes the aforementioned findings and analyses.

A different pooling approach called SPP-net is employed in [2], and the complete network topology is referred to as SPP-net. The main objective of the feature extraction process is to extract a sizable number of features from each proposed region in the image. This proved to be another another crucial argument in favour of convolutional neural network designers including spatial pyramid pooling. SPP is an adaptive method since it can handle images with various scales, sizes, and aspect ratios. Although SPP is discovered to be faster than RCNN, it provides lower accuracy for very deep neural networks.

SPP-net outperforms RCNN in speed, and it demonstrates that MRCNN is simple to create and train. SPP-net not only improves outcomes by accurately estimating various region proposals at their appropriate scales, but it also increases detection effectiveness during the testing phase by dividing the cost of calculation prior to the SPP layer among various proposals. Mask R-CNN is a multitasking method that exhibits systematic faults on overlapping instances and produces bogus edges.

To address this issue, the Mask RCNN adds a branch to predict segmentation masks pixel-by-pixel, parallel to the current branches in Faster R-CNN for classification and bounding box regression. The following advantages were noted in the same paper: Compared to conventional shallow models, CNN has a deeper architecture and offers an exponentially higher expressive power. Additionally, CNN's architecture allows for the combined optimization of numerous related tasks at once.

SPP-net not only improves outcomes by accurately estimating various region proposals at their appropriate scales, but it also increases detection effectiveness during the testing phase by dividing the cost of calculation prior to the SPP layer among various proposals. SPP-net not only improves outcomes by accurately estimating various region proposals at their appropriate scales, but it also increases detection effectiveness during the testing phase by dividing the cost of calculation prior to the SPP layer among various proposals.

It takes more time to train a more complicated and deeper network, but this time consumption can be decreased by putting as many layers into a shared fully convolutional layer. Fast RCNN reduces additional storage costs and enhances accuracy and efficiency. By sharing almost, the entire layer, R-FCN enables the adoption of more potent classification networks to carry out object identification in a fully convolutional architecture. For instance-level recognition, Mask R-CNN provides a versatile and effective framework that may be easily applied to various applications.

In [3], a variety of object identification techniques are covered, including YOLO, Faster RCNN, and SSD. It is demonstrated that Faster RCNN is an improved version of the Fast-CNN. To address the problems of high computation and slow real-time brought on by the selective search strategy in R-CNN and Fast R-CNN, Faster R-CNN uses the region proposal network (RPN). Fast R-CNN uses a method that divides the convolution computation between the region proposal of the input image and the feature layer, greatly reducing the computation. The problems with large computation and slow real-time brought on by the selective search strategy in R-CNN and Fast R-CNN are resolved by Faster RCNN. According to the experiment, test speed for Faster RCNN is 5fp/s, which is 10 times faster than test speed for Fast RCNN. Additionally, the precision has improved. The R-accuracy FCN's is comparable to that of the Faster R-CNN, but its test speed is 2.5 times faster. SSD reduces the complexity of the neural network's computations using the YOLO regression concept to ensure real-time

SSD can extract features from various scales and aspect ratios using the anchor's method to ensure detection accuracy. The paper also emphasises some of the drawbacks, such as the fact that YOLO's detection accuracy has declined somewhat when compared to other cutting-edge deep learning object identification models. The cancel of region suggestion is the main factor contributing to the fall in YOLO's accuracy, which is 66.4% compared to Faster R-accuracy CNN's of 73.2%. The YOLO (You Only Look Once) approach, whose fundamental concept is to divide the image into several grid cells, is highlighted in [4]. The method for classification and localisation is then used on each of these grid cells. The centre of each object in a grid resolves the class label or tag associated with that object.

YOLO forecasts a fixed number of bounding boxes and scores, or class probabilities, for multiple objects for each instance of the class. It is very quick and produces fewer background mistakes. However, it can detect a single object more than once and cannot detect numerous objects within a single grid. With this approach, accuracy rate dropped. RCNN can be

used to find small objects, and YOLO can be implemented in real time. The drawback that RCNN is sluggish and cannot be used is also emphasised in the paper. It mentions how difficult it is for the YOLO technique to locate little objects.
The YOLOv3 technique is detailed in [5]. This method compares different objects by using default kernels on different aspect ratios, and it provides superior precision with a range of sizes. YoloV3 improves performance by using multiscale characteristics for object detection. According to experimental findings, accuracy of 68% has been attained, improving accuracy in comparison to traditional methods. The Dark-Net 53 approach aids in resolving issues with object detection that lack precision.

RNN (Recurrent Neural Network) and LSTM are two deep learning approaches used in [6] to illustrate human fall detection (Long short-term memory). Rectified linear unit (ReLU) and Soft max were selected as the activation function and loss function, respectively, based on the advantages listed below. The model's additional parameters were set to optimise the method's performance.

Based on the receiver operating characteristic (ROC) curve's area under the curve (AUC) statistic, the performance was assessed. Additionally, it compares the model's performance to that of the top depth-map-based fall detection techniques, all of which are constructed using manually created features and have 93% precision and 96% recall at a dropout ratio of 0.5. The performance of the two-layer LSTM was sufficient, and the third layer did not significantly enhance it. This is the first deep learning-based approach to the problem of detecting human falls. It performs better than any other handmade feature-based fall detection algorithms currently in use, with an area under the ROC curve of 0.99.

Human behaviour identification is prioritised in [7], where three deep CNN to channels were utilised to identify human behaviours using weighted hierarchical depth motion maps. The results revealed a 2-9% performance improvement for the majority of the subjects. The multi-task learning methodology shows that it may be used as a hierarchical method to learn a variety of tasks in order to capture intrinsic relationships. The performance of deep learning approaches depends on the size of the network weights, and hyper-parameter adjustment is difficult, among other limitations, which are discussed in the paper.

In [7], the importance of detecting human behaviour is highlighted. Weighted hierarchical depth motion maps were used to identify human actions on three deep CNN channels, and the majority of the subjects showed a 2-9% improvement in performance. The approach to multi-task learning demonstrates its effectiveness as a hierarchical method for learning multiple tasks to capture intrinsic correlations. The paper also discusses some of the drawbacks of deep learning techniques, including the need for large data sets, how network weight size affects performance, how challenging it is to tune hyper parameters, etc.

As a result, it is used in more complex scenarios that not only detect pedestrians but also perform better at identifying human behaviour. A suitable hyper parameter should be carefully picked to fine-tune the model because the results are affected by the hyper parameters employed. The study also makes the case for the need for further development of YOLO algorithms for human detection.

In paper [9], attempts are made to reduce processing time for human face detection and recognition. The system is divided into three stages: motion detection, face detection, and recognition in order to decrease human intervention and boost total system efficiency. Motion detection minimises system processing complexity and search space. The most often used face detection algorithm is the Haar face detector. Although background subtraction takes more time to compute, it significantly shrinks the detection area for faces.

The differences in the query image affect how well the recognition step performs. Every video's opening frame was designed to be the background frame. A moving object was visible in a few videos' background frames. The technique's drawback in these situations was the performance of backdrop modelling.

Discuss Convolutional Neural Network (CNN) methods for identifying human face expressions in [10]. The proposed model's accuracy is between 70.14 and 98.65 percent. The proposed method results in a reduction in computing time, an improvement in validation accuracy, a reduction in loss, and a further performance evaluation that contrasts the current model with the prior one. All test photos from the dataset were used to evaluate performance. The proposed model produced an accuracy of 70.14 percent on average considering emotions of neutral, sad, emotions, fear, disgust.

Discuss Convolutional Neural Network (CNN) methods for identifying human face expressions in [10]. The proposed model's accuracy is between 70.14 and 98.65 percent. The proposed method results in a reduction in computing time, an improvement in validation accuracy, a reduction in loss, and a further performance evaluation that contrasts the current model with the prior one. All test photos from the dataset were used to evaluate performance, and the model scored 95 percent for happy, 75 percent for neutral, 69 percent for sad, 68 percent for surprise, 63 percent for disgust, 65 percent for fear, and 56 percent for furious. The proposed model produced an accuracy of 70.14 percent on average.

[11] discusses the use of deep learning for human face and spoofing detection. The CNN technique is used to determine with accuracy whether or not there is a real human face in the frame of the detecting camera. A validation accuracy of 93% and a training accuracy of 95% are reached after 30 training epochs. On the test set, which included 3300 facial images and 3300 facial images wearing 3D masks, the suggested CNN performed exceptionally well with an accuracy rate of 86.75%. With "No Face" as the relevant class, the test set's recall, precision, and F1-score are, respectively, 0.8, 0.93, and 0.86.

Real-time object detection in transportation: According to [12], object detection methodologies have various limitations, one of which being YOLOv3's inability to recognise objects in real-time. To address this problem, MobileNetv1

yolov3lite is proposed. The model was modified to achieve real-time object identification by using MobileNetv1 instead of Darknet53 and yolov3lite to enable feature fusion.

Utilizing optimization techniques has allowed for a nearly 14-fold reduction in computation complexity. However, altering the training procedures and loss function can enhance accuracy losses. Buses, bicycles, pedestrians, and car images are used as datasets, and the model detects and categorises them with a respectable level of accuracy. Real-time object detection for improved energy effectiveness

A description of object identification techniques to use for greater power efficiency can be found in [13]. CNN is a popular technique for computer vision tasks, but its capabilities are limited by its high computational and storage costs. GPU cannot be used in real-time mobile applications because of its high power consumption. Here, SVD and low bit data format are used for algorithm decomposition to reduce algorithm complexity. The Wino grad algorithm is used to encourage the hardware's peak performance.

In [13], SSD and the YOLO detection method were able to achieve a power efficiency of 43.5% and an accuracy of 73%. The utilisation of hardware platforms can be improved by pipeline design in heterogeneous systems and memory optimization techniques like cross layer strategy. The detection algorithm framework can be improved with development to increase accuracy and reduce energy consumption.

Research on deep learning-based computer vision technology is conducted in the study [14], intelligence and automation of monitoring systems issues can be addressed by deep learning technology. In accordance with the three-dimensional recognition of the human body, a multi view fusion model based on attention mechanism is proposed together with a set of deep learning monitoring systems. A system can monitor a scene, track and identify persons there, and evaluate the data it gathers to perform oversight, early warning, and preventative functions.

Deep learning monitoring system research therefore has significant and vital application value. In order to research the problem of 3D human action recognition, this work proposes an attention-based multi view re observation fusion model. The attention mechanism is employed throughout the fusion process to assess if the observation perspective supports action recognition based on action sequence data. . By using action information, the model can learn to select among a variety of viewpoints those that are appropriate for action recognition.

[15] Illustrates how innovation and change have been sparked by recent developments in augmented reality and artificial intelligence. For augmented reality (AR) systems to function optimally, the real environment and its elements must be precisely identified. By scanning the environment in real-time, the majority of AR technologies can identify a 3D spatial map of the parts.

This paper also explains how object detection has been used extensively in augmented reality. Various types of algorithms can be used in augmented reality applications, depending on the amount of data that is available, how the AR device is set up, and the objectives of the detection. Marker-based techniques and statistical classifiers are the main tools used in traditional AR object detection, which greatly aids in object detection. The Table 1 gives the summary of all deep learning methods and comparison of the same.

Table 1. Comparison of Object Detection methods

No.	Object Detection Techniques	Advantages	Limitations
1	Sliding Window	Easy to implement and simple	The method is time consuming
2	RCNN	Number of regions proposed is less as compared with sliding window technique	Expensive training with respect to time and space. object detection is slow
3	Overfeat	Higher speed when compared RCNN	Accuracy is less
4	SPP-net	Faster than R-CNN Avoid repeated computation of features.	Accuracy is less for deep neural network
5	MRCNN	Easy to train and a small overhead	Unsuitable for real time applications
6	Attention Net	Accuracy of detection of objects is higher	Has a low recall
7	Fast R-CNN	Ensures high quality detection	Slow clustering and high computation time
8	Faster R-CNN	As the name indicates, it is faster when compared to Fast R-CNN	Implementation is slower than YOLO
9	DeepIDNet	Deformation of objects learning	Encounter issues on verification of detection

10	SSD	Suitable for bigger objects	Cannot generate higher level features to small objects
11	YOLO	Highly fast and less background errors compared to Fast R-CNN	Possibility of detecting same object multiple times, Loss in accuracy rate and cannot detect multiple objects in the same grid
12	RFCN	Easier to train with less complexity and faster than	Require large amount of computation sources
13	FPN	Has rich semantics in all the levels	Reduced accuracy
14	DeNet	Anchors which are predefined are not required and it is faster than RCNN	Time consuming in generating corners and evaluation of base network

III. CONCLUSION

In this survey paper various object and human detection methods have been analysed by underscoring the importance, merits, demerits, performance and limitations. This work also provides an overall view of the approach, which the researchers have tried to implement for the models proposed. The future scope of the detection methods is also considered in this paper with the intention of supporting innovation to drive diverse applications for the benefit of mankind.

References

[1]. Bhagya C, Prof.Shyna, "An Overview of Deep Learning Based Object Detection Techniques", Conference in innovation and communication Technology, 2019.

[2]. Zhong-QiuZhao , Peng Zheng, Shou-Tao Xu, and Xindong Wu, "Object Detection with Deep learning : A Review", IEEE transactions on Neural Networks and Learning Systems, 2019.

[3]. Cong Tang, Yunsong Feng, Xing Yang Chao Zheng, Yuanpu Zhou, "The Object Detection Based on Deep Learning", 4th International Conference on Information Science and Control Engineering, 2017.

[4]. Priyanka Malhotra,Ekansh Garg, "Object Detection Techniques: A Comparison",IEEE 7th International Conference on Smart Structures and Systems ICSSS, 2020.

[5]. William Tarimo, MoustafaM Sabra, ShonaHendre,"Real-TimeDeep Learning-Based object detection", IEEE Symposium Series on Computational Intelligence (SSCI), 2020.

[6]. AnahitaShojaei- Hashemi1, Panos Nasiopoulos1, James J. Little2 and Mahsa T. Pourazad, "Video-based Human Fall Detection in Smart Homes", IEEE International Symposium on Circuits and Systems, 2018.

[7]. Di Wu, Nabin Sharma, and Michael Blumenstein," Recent Advances in Video-Based Human Action Recognition using Deep Learning: A Review", International Joint Conference on Neural Networks, 2017.

[8]. Jia Lu, Wei Qi Yan and Minh Nguyen, "Human Behaviour Recognition Using Deep Learning", IEEE Conference on Advanced Video and Signal based Surveillance, 2018.

[9]. S. V. Tathe, A. S. Narote, S. P. Narote, "Human Face Detection and Recognition in Videos", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016.

[10]. AkritiJaiswal, A. Krishnama Raju, Suman Deb, "Facial Emotion Detection Using Deep Learning", International Conference on Emerging Technology, 2020.

[11]. Saankhya Mondal," Implementation of Human Face and Spoofing Detection Using Deep Learning on Embedded Hardware", International conference on Computing and networking technology, 2020.

[12]. Na Wang, Ruoyan CHEN, Kang Xu, "A Real-time Object Detection Solution and Its Application in Transportation", IEEE 3rd International Conference on Communications, Information System and Computer Engineering, 2021.

[13]. Yincheng Yu1, Kaiyuan Guo1, Yiming Hu1, Xuefei Ning1, Jiantao Qiu1, Huizi Mao1, Song Yao2, Tianqi Tang1, Boxun Li1, Yu Wang1, and Huazhong Yang1, "Real-time object detection towards high power efficiency", Design, Automation and Test in Europe Conference and Exhibition, 2018.

[14]. Zhipan Wu and HuayingDu, "Research on Human Action Feature Detection and Recognition Algorithm Based on Deep Learning", It is an open access article distributed under the Creative Commons Attribution License, 2022.

[15]. Yalda Gasemi, Heejin Jeong, sunghochoi, Kyeong- Beom Park, Jae YeolLee, "Deep learning-based object detection in augmented reality: A systematic review", this research was supported by the Republic of Korea's MSIT, 2022.