E-ISSN: 2664-8784
P-ISSN: 2664-8776
Impact Factor: RJIF 8.26
IJRE 2025; 7(2): 11-15

www.engineeringpaper.net
Received: 12-04-2025
Accepted: 17-05-2025

Dr. Anju J Prakash
Associate Professor, Department of Computer Engineering, Amal Jyothi College of Engineering, Kanjirappally, Kerala, India

Correspondence
Dr. Anju J Prakash
Associate Professor, Department of Computer Engineering, Amal Jyothi College of Engineering, Kanjirappally, Kerala, India

Enhanced object detection in video streams using deep learning and RCNN Architectures

Anju J Prakash

DOI: https://www.doi.org/10.33545/26648776.2025.v7.i2a.108

Abstract
As a consequence of the digital era and, more specifically, videos, such as television archiving and video surveillance, a tremendous quantity of data is created each and every day. This is particularly true of videos. If we want to keep control over this material and make it accessible for analysis, categorization, and a number of other applications, it is evident that we will need algorithms that are capable of doing this work in a timely and efficient manner. The proposed approach makes it possible to do an analysis of video clips by making use of deep learning methods. The main objective of this study is to design an object detection architecture that is more precise than existing methods. Object detection is a method that can recognize and detect a variety of elements that are visible in an image or video and label them in order to categorize these objects. This may be accomplished via the use of computer software. Deep learning object recognition is a technique that is both rapid and accurate in its ability to anticipate the location of an item inside a picture. This technology has the potential to be beneficial in a number of contexts. RCNN is one of the innovative approaches that may be utilized in conjunction with deep learning to detect objects.

Keywords: Object detection, Deep learning, RCNN

Introduction
The primary objective of this study is to provide a more precise architecture for object detection. There has been an increase in the accessibility of video data due to the fact that it often contains more information presented in a more straightforward manner. This is to be anticipated in many application domains, but the advantages of the approaches used to exploit it are limited since they still need human supervision. Scanners, cameras, and improved storage systems are other key factors. In this specific scenario, the multimedia community has been tasked with exploring and developing more effective methods of video analysis, including but not limited to classification and the identification of activities. Retrieving or summarizing films for easy browsing is becoming more dependent on these methods. Exploitation of large video databases is a rapidly expanding area of research. Published publications feature algorithms or tools developed with advanced methods of machine learning.

Deep learning techniques includes
1.Object detection
Object detection, in its most basic form, is a method used to locate and catalog the many things seen in a picture or video for the purpose of categorization. Algorithms that employ deep learning to deliver effective results are widely used for object identification in order to recognize and identify things using a variety of ways.
One of the many applications of deep learning object identification is the prediction of an item's location in an image. RCNN, or Region-based Convolutional Neural Networks, is one of the cutting-edge techniques utilized in object recognition using deep learning.

2. Extraction of features
Feature extraction is a part of the dimensionality reduction strategy, which reduces the size of a large dataset by breaking it down into smaller, more manageable chunks. This will make processing much easier. The most important aspect of these massive data sets is the abundance of variables they include. It takes a lot of computational resources to process these variables. By selecting and merging variables into features, feature extraction aids in

Standardization of Wearable and EHR Data

extracting the best feature from such massive data sets. These characteristics have a low learning curve and a high ability to effectively characterize the underlying data.

Feature extraction is useful in situations when a huge data collection must be efficiently processed while avoiding the loss of any essential details. Using feature extraction, we were able to reduce the quantity of superfluous information in our data set. Finally, data minimization boosts learning speed and generalization stages of machine learning, while also allowing the model to be constructed with little computational overhead.

3. A network for classification

Classification is an operation that may be performed on data that is both organized and unstructured.

The act of dividing a given collection of data into distinct categories is known as classification. The technique begins with the first stage, which is to make an educated guess as to the category that the data points in question belong to. It is common practice to refer to the classes using the phrases target, label, and classes to describe them. Classification predictive modeling refers to the process of trying to come up with an approximation of the mapping function that goes from discrete input variables to output variables. The primary purpose consists of determining which category or category set the new data will fall within.

Literature Review

ResNet50

The name "ResNet-50" refers to a convolutional neural network that has fifty layers. A version of the network that has been pretrained is included in the ImageNet database. This version of the network was trained using more than one million images. The trained network is able to recognize photographs from one thousand distinct item categories, including a wide variety of animals, a keyboard, a mouse, and many other things. As a result, the network has amassed a vast collection of feature representations for a number of different pictures. Images with a resolution of 224 by 224 pixels may be uploaded to the network without issue. A modification of the first ResNet design, known as ResNet-34, had a total of 34 weighted layers. The VGG neural networks (VGG-16 and VGG-19) served as the inspiration for the construction of each individual convolutional network that makes up the regular network. A ResNet, on the other hand, is not as complicated as a VGGNet and has a smaller number of filter nodes. When compared to the 19.6 billion FLOPs that can be accomplished by a VGG-19 Network, a ResNet with 34 layers can perform at 3.6 billion FLOPs, while a ResNet with 18 layers can achieve 1.8 billion FLOPs.

R-CNN

In 2014, a group of researchers at UC Berkeley developed a deep convolutional network known as the R-CNN (region-based convolutional neural network). This network has the ability to recognize eighty distinct types of objects that are shown in photographs. The most important contribution made by R-CNN is the straightforward extraction of features from a convolutional neural network (CNN).

In order to address a number of deficiencies inherent in the R-CNN model, the Fast R-CNN model has been proposed as a potential upgrade. Ross Girshick, now an AI researcher at Facebook and formerly employed as a researcher at Microsoft, is the sole creator of the object detector known as Fast R-CNN. Fast R-CNN is an effective solution for a number of R-CNN issues. As its name implies, the Fast R-CNN is far faster than the regular R-CNN.

Faster R-CNN

The Faster R-CNN object detection model outperforms the Fast R-CNN model thanks to the addition of a region proposal network (RPN) to the CNN model. Both the RPN and the detection network share full-image convolutional features, which makes it possible to provide almost free suggestions for regions. It is a fully convolutional network that makes predictions about object boundaries and objectness ratings everywhere and all at once. The RPN has received comprehensive training and is now capable of producing first-rate region suggestions, which Fast R-CNN employs for detection.

RPN and Fast R-CNN are able to be integrated into a single network thanks to the sharing of their convolutional features, with the RPN component of the combined network leading the search.

Methodology

The proposed system consists of two phases: Implementation phase and training phase.

The implementation phase consists of

Object Detection

Object detection begins with the extraction and preprocessing of individual frames from a video. Image preprocessing often involves scaling and normalization. The frame is then sent to a model for feature extraction known as ResNet50.

ResNet50

ResNet-50 is a 50-layer convolutional neural network. A version of the network that has already been trained on over a million images is available in the ImageNet database. Among the 1000 item classes that the trained network can recognize are several animals, a computer keyboard, a mouse, and many more. As a result, the network has learned to represent many different types of pictures using a wealth of features. Images of up to 224 by 224 pixels are supported by the network.

There are two main tenets of design that ResNet adheres to. First, the number of filters in each layer remains constant regardless of the size of the output feature map. Second, there are twice as many filters as there are layers to ensure the same temporal complexity regardless of how small the feature map is.

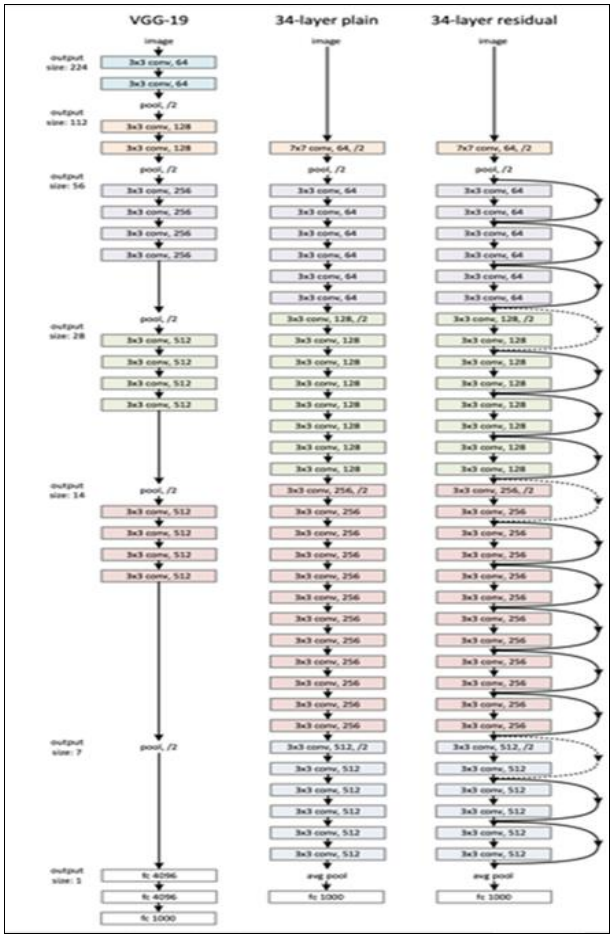

Fig 1 ResNet model architecture

The input picture is subjected to convolution and pooling, producing an image in either two or three dimensions. An object detection model then gets the data.

R-CNN

In 2014, a group of UC Berkeley researchers developed a deep convolutional network called the R-CNN (region-based convolutional neural network) that can identify eighty distinct types of objects in photos. R-CNN's key contribution is the ease with which it extracts features from a CNN.

To address the limitations of the R-CNN paradigm, the Fast R-CNN model is proposed. Facebook AI researcher and former Microsoft Researcher Ross Girshick created the object detector Fast R-CNN. Fast R-CNN provides answers to a number of issues with R-CNN. The Fast R-CNN, as the name implies, is faster than the regular R-CNN.

While the Fast R-CNN model does have its advantages, the fact that it uses the laborious Selective Search technique to generate area suggestions is a major drawback. The development of Faster R-CNN has allowed for the correction of this shortcoming. Faster R-CNN uses a region proposal network (RPN) to improve upon Fast R-CNN's speed. In this study, we present our proposal for a faster version of R-CNN.

Faster R-CNN

Faster R-CNN is a more effective object detection model than Fast R-CNN because it combines a region proposal network (RPN) with the CNN model. Network for detection and the RPN trade full-image convolutional features, allowing for almost cost-free area suggestions. It is a fully convolutional network that simultaneously predicts global object bounds and objectness scores. Fast R-CNN relies on RPN's top-notch region recommendations, which are the result of extensive training. RPN and Fast R-CNN may be integrated into a single network thanks to their shared convolutional features, with the RPN component guiding the combined network's search.

This model consists of

1.The region proposal network (RPN) is a fully convolutional network that generates proposals at various sizes and aspect ratios. The RPN uses neural network jargon to specify an area of interest for the object identification (Fast R-CNN).

2. In lieu of pyramids of photographs (i.e., multiple instances of the image at varying scales) or pyramids of filters (i.e., multiple filters with variable sizes), anchor boxes were presented in this work. Anchor boxes are used as a point of reference and are defined by their size and ratio. Multiple reference anchor boxes in the same area means its dimensions and aspect ratio might vary. You may see this as a reference anchor box pyramid. Then, items of varying sizes and shapes may be located by first mapping each area to its own unique reference anchor box.

3. Convolutional calculations are used by both the RPN and the Fast R-CNN. This leads to a reduction in processing time.

It consists of 2 modules

***RPN*:** To produce proposals for regions.

Rapid R-CNN: For finding things in the suggested areas.

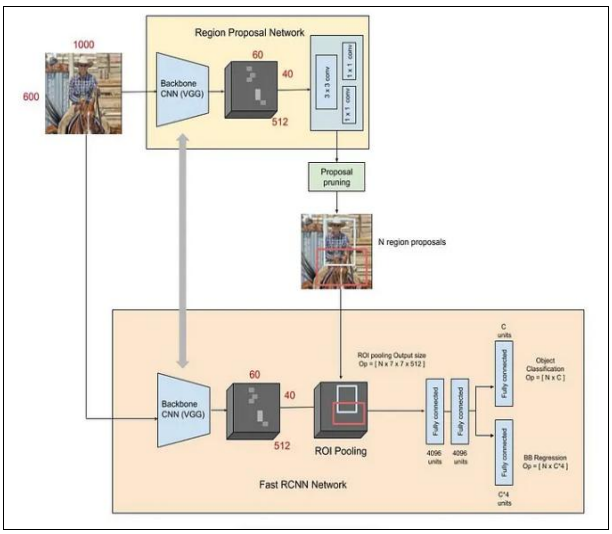

Fig 2: The architecture of Faster R-CNN

The model consists of two portions

1. Region Proposal Network
2. Classification Network.

Region Proposal Network

The object-bound and objectless scores for each place are predicted using a fully convolutional network termed the Region Proposal Network (RPN). After being trained, the RPN generates highly accurate region suggestions that Fast R-CNN utilizes to make its detections. Its purpose is to provide suggestions on what may be in a picture based on its content.

To now, RPN has shown to be highly good at object detection using R-CNN, making it the only true backbone. Its purpose is to provide suggestions on what may be in a picture based on its content.

We have learnt how to use Selective Search to retrieve many areas from R-CNN. The disadvantages of the Selective Search include being time-consuming to calculate and an offline method. At this juncture, the region proposal network becomes relevant. The Region Proposal Network, created by Quicker R-CNN, employs a modest network to propose regions. A classifier built into RPN calculates the likelihood of a given area. The regressor also provides the bounding box coordinates.

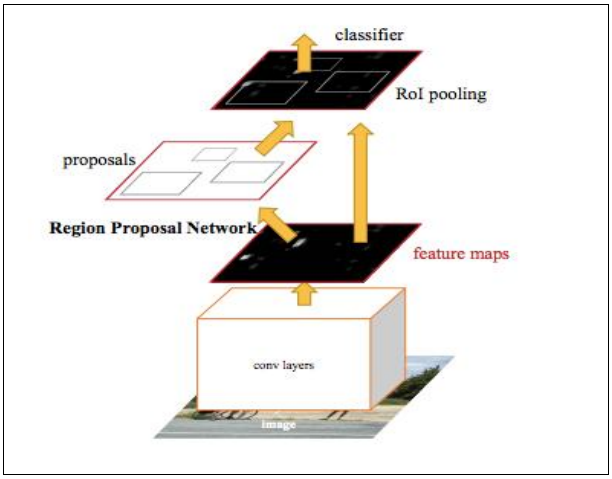

Fig 3: Architecture of Regional Proposal Network

Classification Network

To identify the precise category to which an item belongs, we use a classification network. Classification is performed when a number of characteristics from the picture are extracted and fed into the RPN. The resulting information is then sent on to the tracking mechanism. The classification network will choose the most valuable category based on the output. Classification neural networks gain tremendous power when used in tandem with a variety of prediction neural network types in a hybrid system. In systems with numerous unique nonlinear operating areas, the classification neural network may determine the process operating area before going on to the appropriate prediction neural network.

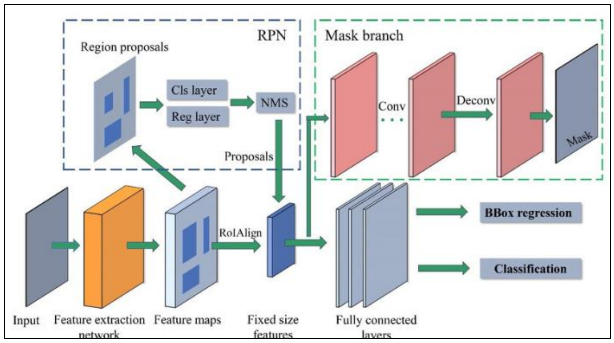

Fig 4: Classification Network

The training phase consists of

1. Compilation of datasets
2. Labelling or annotation
3. Training the model

Compilation of datasets

I was able to recall the images of the objectives that needed to be accomplished. A machine learning dataset is a collection of data that is used to train a machine learning model. An example is given to the deep learning algorithm in the form of a dataset so that it may learn how to make predictions.

Labelling

An item in the picture may be given a name by first drawing a square box around it and then storing the name of the box along with the picture itself.

A technique for machine learning known as dataset labelling enables the identification of raw data and the labeling of data that is useful and relevant for the purpose of providing context for the raw data. After then, the data might be used by machine learning to learn from it. The process of data labeling is an essential step that must be taken in order to increase both the scalability factor and the quality factor. This is because data labeling may provide context to data before that data is used in the training model. For instance, if we have a photograph, labeling may tell us if the picture depicts an animal or an automobile, and the corresponding word may appear in the audio recording. This also happens if we have an x-ray report that pertains to a person's medical history. The internal approach, the outsourced approach, the crowd-sourcing approach, and the machine approach are some of the approaches that are included. Labeling datasets may be accomplished via the use of a wide range of approaches, or through the combination of a number of different methods.

Preparing or Training the model

Training to recognize items is carried out by providing them with information in the form of a picture and annotation. Training a machine learning algorithm requires the use of a dataset that is referred to as a training model. It is made up of many sets of pertinent inputs and acquire knowledge of the ideal values for each of the pertinent qualities. The two types of machine learning models that are used the most often are supervised and unsupervised learning. There are many more forms of machine learning models.

Modelling

In the process of object detection, the initial frame of a video is taken, and the frames are then subjected to pre-processing. The pre-processing stage involves activities such as scaling and normalization. After that, the frame that has been retrieved is sent to a model for feature extraction known as ResNet50. The Region Proposal Network, often known as RPN, is a fully convolutional network that makes predictions for both object-bound and objectless scores at each point. During training, the RPN is taught to provide high-quality region suggestions, which the Fast R-CNN later employs for detection. Its purpose is to point out several recognizable objects hidden inside a picture that has been provided. For the purpose of identifying the precise category to which an item belongs, we make use of a network that categorizes things. Classification is performed when a number of characteristics from the picture that are recognizable by the RPN have been extracted. Following this, the tracking component is given the result that was acquired. The classification network chooses whatever category to select based on which has the highest output value.

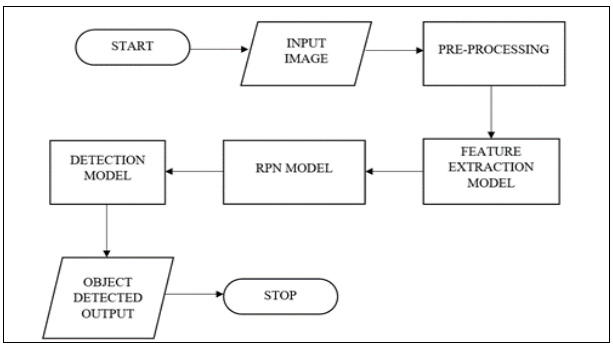

Fig 4: Basic flow diagram

Results and Discussion

The result of the detection process consists of the item that is being categorized and its position in the picture. To do this, bounding boxes are drawn around the item in the picture. The output of an object detection model is composed of the following three components: Using the COCO file format, the bounding boxes are represented as x1, y1, width, and height. The kind of the box that is being bound. The probability score for that prediction, or the degree to which the model is certain that the class in question is, in fact, the class that was predicted for it.

Conclusion

A model for more accurate object identification is built here, and it's based on R-CNN. By using the COCO dataset, which is a pre-trained dataset, it is possible to find objects and generate bounding boxes around them. Because of this paradigm, it is now able to recognize a variety of components within a single image. By merging an area with a standard R-CNN, the object detection model known as Faster R-CNN is able to surpass its predecessor, Fast R-CNN.

The RPN and the detection network are able to offer virtually free area suggestions via the exchange of full-image convolutional features. The network is entirely convolutional and makes predictions about object limits and objectness ratings concurrently everywhere. Accurate detections can be achieved with the help of Fast R-CNN thanks to the RPN's completely trained region suggestions. RPN and Fast R-CNN are blended into a single network by sharing the convolutional features of both of their previous networks; the RPN component instructs the combined network on where to look for results. The fact that object detection is carried out in a manner that is more accurate compared to other technologies that are used in the detection process is the primary benefit of this system. An picture may be used to identify a number of different items.

In spite of the fact that we have employed a model for object recognition that has a higher level of accuracy and that can recognize numerous things inside the same picture, this model is not suitable for use in recognizing objects within movies. Therefore, in the future we need to concentrate on overcoming this drawback, and we need a better result.

References

1. Nahuel E. Albayrak, “Object Recognition using TensorFlow,”2020.
2. Rasika Phadnis, Jaya Mishra, Shruti Bendale, “Objects Talk - Object detection and Pattern Tracking using TensorFlow,” 2020
3. Jemai Bornia, Ali Frihida, Christophe Claramunt, “Detecting objects and people and tracking movements in a video using TensorFlow and deeplearning,”2020
4. Wang Yang, Zheng Jiachun, “Real-time face detection based on YOLO”,
5. Abhishek Sarda, Dr. Shubhra Dixit, Dr. Anupama Bhan “Object Detection for Autonomous Driving using YOLO algorithm,”2021
6. https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e
7. https://medium.com/egen/region-proposal-network-rpn-backbone-of-faster-r-cnn-4a744a38d7f9