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, Issue 5, April 2023

Survey Paper on Real-Time Object Detection

Pratik J. Rangari[1], Chandrakant Pandit[2], Pallavi A. Tayade[3], Mahesh P. Konde[4], Shivani Sable[5], Prof. Dipali A. Sananse[6]
U.G. Students, Department of Computer Science and Engineering[1,2,3,4,5]
Assistant Professor, Department of Computer Science and Engineering[6]
Jawaharlal Darda Institute of Engineering and Technology, Yavatmal, Maharashtra, India

Abstract: *Object detection is one of the most important and challenging branches of computer vision, which has been widely applied in people's lives, such as monitoring security, autonomous driving, and so on, to locate instances of semantic objects of a certain class. Conventional object detection algorithms were primarily derived from machine learning. This involved the design of features for describing the object's characteristics followed by integration with classifiers. In recent years, the application of deep learning (DL), and more specifically Convolutional Neural Networks (CNN) has elicited great advancement and promising progress and has, therefore, received much attention on the global stage of research about computer vision.*

Keywords: Real Time, Machine Learning, Object Detection, Neural Networks, RCNN, SSD, Caffe model, COCO Dataset.

I. INTRODUCTION

The information age has witnessed the rapid development of wireless network technology, which has attracted the attention of researchers and practitioners due to its unique characteristics such as flexible structure and efficiency. As wireless network technology continues to evolve, it has brought great convenience to people's life and work with its powerful technical capabilities. Wireless networks have gradually facilitated the mainstream of people's online life. At the same time, the advent of the 5G network will further enable the greater development and more advanced applications of wireless network technology. The future generations of wireless networks will provide strong support for related applications. Many of these applications connect and transmit information within networks based on the detection of specific target objects. To achieve a comprehensive network connection between people and people, things and people, and things and things, one of the key tasks of future applications is to identify the target in a real-time manner in the wireless
network's platform for the city guide and can search place in the city without taking the help of any personal guide.[11]

The study of "CV," or CV for short, aims to develop methods that will allow computers to "see" and understand the content of computerized images such as pictures and videos. Because people, particularly children, illuminate the problem of CV insignificantly, it appears to be a simple one. It is, by the way, a generally unresolved issue due to both the limited understanding of natural vision and the complexities of vision discernment in an active and constantly changing physical world. The significance of a CV lies in the issues it can shed light on. It is one of the most cutting-edge technologies, allowing communication between the developed and developing worlds. CV allows self-driving cars to understand their surroundings. Cameras in various locations around the vehicle record video and feed it to a CV program, which creates images in real-time to identify activity signs and street limits.[1]

The study's main contribution is the design and implementation of real-time object detection and recognition systems using the SSD algorithm and deep learning techniques with a pre-trained model. Our proposed system can detect static and moving objects in real-time and classify them. The primary goals of this study were to investigate and develop a real-time object detection system that uses deep learning and neural systems to detect and recognize objects in real-time. Furthermore, we tested the free, pre-trained models with the SSD algorithm on various types of datasets to determine which models have high accuracy and speed when detecting an object. Besides this, the system must be operable on reasonable equipment. During the coding procedure, we evaluated various deep learning structures and techniques and developed and proposed a highly accurate and efficient object detection system.[9]

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, Issue 5, April 2023

II. METHODS OF OBJECT DETECTION

This paper has been completed based on the information gathered from various sources. Internet, various research papers, and articles have been surveyed to gather and understand the information given below.

This section presents a proposed approach for detecting objects in real-time from images by using a convolutional neural network deep learning process. Some algorithms such as CNN, faster RCNN, YOLO, and SSD are only suitable for highly powerful computing machines and they require a large amount of time to train.

2.1 CNN

CNN stands for Convolutional Neural Network. It is a type of deep neural network that is commonly used for image and video analysis tasks such as image recognition, object detection, and segmentation

CNNs are designed to process data with a grid-like structure, such as a 2D image. They consist of multiple layers of interconnected nodes, with each layer learning and extracting different features from the input image.

The first layer of a CNN is typically a convolutional layer, which applies a set of filters to the input image to extract specific features. These features can include edges, lines, and other patterns. The output of the convolutional layer is then passed through a non-linear activation function, such as the rectified linear unit (ReLU), to introduce non-linearity into the model.

After the convolutional layers, the output is typically passed through a pooling layer, which reduces the spatial dimensions of the feature maps while preserving the important features. This helps to reduce the computational complexity of the model and prevent overfitting.[2]

Finally, the output of the pooling layer is passed through one or more fully connected layers, which combine the features to make a prediction. The output of the last fully connected layer is typically passed through a SoftMax function to produce a probability distribution over the possible classes.

CNNs have been very successful in the image and video analysis tasks and have achieved state-of-the-art performance in many benchmark datasets. They are a key component of many real-world applications, including self-driving cars, facial recognition systems, and medical image analysis.

2.2 Faster RCNN

I believe you might be referring to the Faster R-CNN (Region-based Convolutional Neural Network), which is a popular object detection algorithm that builds on the original R-CNN method.

The Faster R-CNN method is designed to improve the speed and accuracy of the R-CNN method. It does this by introducing a Region Proposal Network (RPN) that generates object proposals in a more efficient way than the selective search algorithm used in R-CNN. The RPN shares convolutional features with the object detection network and generates object proposals by sliding a small window over the convolutional features map and predicting abjectness scores and bounding box offsets at each position.

The object proposals generated by the RPN are then used as input to the object detection network, which predicts the class of the object and refines the bounding box coordinates. The object detection network is typically a Fast R-CNN network, which uses the RoI (region of interest) pooling layer to extract features from the region proposals.[3]

The Faster R-CNN method has achieved state-of-the-art performance in many benchmark datasets and is widely used in real-world applications, such as autonomous driving, surveillance systems, and medical image analysis.[2]

2.3 YOLO

YOLO (You Only Look Once) is a popular object detection algorithm that was introduced in 2016 by Joseph Redmon et al. The YOLO algorithm is a type of deep neural network that performs object detection by directly predicting the bounding boxes and class probabilities of objects in an image, all in one pass through the network.

In YOLO, the input image is divided into a grid of cells, and each cell is responsible for detecting objects that fall within that cell. The output of the network consists of a set of bounding boxes and class probabilities for each cell. Each bounding box consists of four values: x, y, width, and height, which represent the centre point and size of the object.

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, Issue 5, April 2023

To generate these outputs, YOLO uses a deep convolutional neural network (CNN) that is trained on labelled images. The network consists of 24 convolutional layers followed by 2 fully connected layers. The output of the final layer is a tensor that contains the bounding box coordinates and class probabilities for each cell in the image.[10]

YOLO also uses a loss function that combines both localization and classification losses to optimize the network. The localization loss measures the difference between the predicted and ground-truth bounding boxes, while the classification loss measures the difference between the predicted and ground-truth class probabilities.

One of the main advantages of YOLO over other object detection algorithms is its speed. Because it only requires a single pass through the network to detect objects in an image, it is much faster than other methods that require multiple passes. This makes it well-suited for real-time applications such as autonomous driving, robotics, and surveillance systems.[6]

In summary, YOLO is a deep neural network that performs object detection by directly predicting the bounding boxes and class probabilities of objects in an image in one pass through the network. It is fast, accurate, and has achieved state-of-the-art performance on several benchmark datasets.

2.4 SSD

SSD (Single Shot MultiBox Detector) is a popular object detection algorithm that was introduced in 2016 by Wei Liu et al. SSD is a type of deep neural network that performs object detection by predicting the bounding boxes and class probabilities of objects in an image using a single feedforward pass through the network.

Like YOLO, SSD is a fully convolutional neural network that processes the entire image in a single pass. The input image is passed through a series of convolutional layers that extract feature maps at multiple resolutions. These feature maps are then used to generate a set of default bounding boxes at different aspect ratios and scales for each spatial location in the feature maps. The default bounding boxes are then refined based on the predicted bounding box offsets and class probabilities.

SSD uses a loss function that combines both localization and classification losses to optimize the network. The localization loss measures the difference between the predicted and ground-truth bounding boxes, while the classification loss measures the difference between the predicted and ground-truth class probabilities. SSD also uses hard negative mining to select a subset of negative examples that are the most difficult to classify, which helps to improve the accuracy of the model.[9]

One of the main advantages of SSD over other object detection algorithms is its ability to detect objects at multiple scales and aspect ratios. Because SSD uses feature maps at multiple resolutions, it can detect small objects as well as large objects in the same image. This makes it well-suited for applications such as robotics, autonomous driving, and surveillance systems.[8]

In summary, SSD is a fully convolutional neural network that performs object detection by predicting the bounding boxes and class probabilities of objects in an image using a single feedforward pass through the network. It is fast, accurate, and has achieved state-of-the-art performance on several benchmark datasets.

2.5 MobileNet

MobileNet is a lightweight CNN architecture with limited computational resources for mobile devices. It is efficient and can be used for real-time object detection on mobile devices.

2.6 Tiny YOLO

Tiny YOLO is a smaller version of the YOLO architecture that is designed to run on less powerful devices. It is optimized for speed and can achieve real-time object detection on devices with limited computational resources.

III. RESULTS AND DISCUSSION

Generally, the output of real-time object detection is a set of bounding boxes that indicate the location and size of each detected object in the input image or video frame. The bounding boxes are often accompanied by a label that identifies the type of object detected (e.g., person, car, tree, etc.).

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, Issue 5, April 2023

The accuracy and speed of real-time object detection can vary depending on the complexity of the objects being detected, the quality of the input data, the computational resources available, and the specific algorithms used. In general, the goal is to achieve high accuracy while maintaining real-time performance, which typically means processing input data at a rate of at least 30 frames per second.

3.1 What is Machine Learning (ML)

Machine learning is a field of artificial intelligence (AI) that focuses on building computer algorithms that can learn from data and make predictions or decisions without being explicitly programmed. The goal of machine learning is to enable machines to learn and improve their performance over time based on the data they receive.

The concept of machine learning has its roots in the field of artificial intelligence, which began in the 1950s. The term "machine learning" was coined in 1959 by Arthur Samuel, who is considered one of the pioneers of the field. Samuel was a computer scientist who worked at IBM, and he is credited with developing one of the first self-learning programs, a checkers-playing program that improved its performance over time through experience.

Since then, the field of machine learning has grown rapidly, and many researchers and practitioners have made significant contributions. In the 1990s, machine learning algorithms began to be used in commercial applications such as spam filtering and fraud detection. In the early 2000s, machine learning was used in speech recognition and natural language processing, and it has since been applied to a wide range of fields, including computer vision, robotics, healthcare, and finance.

One of the key breakthroughs in machine learning came in 2012 when a deep learning algorithm developed by researchers at the University of Toronto won a competition in image recognition. The algorithm, called Alex Net, was able to significantly improve the accuracy of image recognition and sparked a renewed interest in deep learning.

Today, machine learning is a rapidly growing field with many applications and opportunities for research and development. With advances in computing power and data availability, machine learning is becoming increasingly important in many industries and fields, and is expected to continue to have a significant impact in the years to come.

3.2 What is Neural Network

Neural networks are a type of machine learning algorithm that is inspired by the structure and function of the human brain. They are composed of layers of interconnected nodes or neurons that process input data and produce output predictions or classifications.

Neural networks can be trained on large datasets to learn complex patterns and relationships in the data. During training, the network adjusts the weights and biases of its neurons to minimize the difference between the predicted outputs and the actual outputs. This process, called backpropagation, allows the network to improve its accuracy over time.

3.3 What is Caffe Model

Caffe is a deep learning framework that is often used for image recognition tasks. It was developed by the Berkeley Vision and Learning Centre (BVLC) at the University of California, Berkeley, and is now maintained by the open-source community.

A Caffe model is a deep neural network that has been trained on a specific task, such as image classification or object detection. The model is composed of layers of neurons that process the input data and produce output predictions. These layers can include convolutional layers, pooling layers, fully connected layers, and more.

To create a Caffe model, a user typically starts by defining the architecture of the network, including the number and type of layers. The user then trains the model on a large dataset using backpropagation and gradient descent, adjusting the weights and biases of the neurons to minimize the difference between the predicted outputs and the actual outputs.

Once the model is trained, it can be used for inference on new data, making predictions or classifications based on the patterns it has learned during training. Caffe models have been used in a wide range of applications, including image recognition, object detection, and segmentation, and have achieved state-of-the-art performance on many benchmarks.

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, Issue 5, April 2023

3.4 What is COCO Dataset

The COCO (Common Objects in Context) dataset is a large-scale image recognition, segmentation, and captioning dataset that is commonly used for object detection and recognition tasks. The dataset was developed by Microsoft

The COCO dataset consists of more than 330,000 images, each of which is labeled with object category annotations, object segmentation masks, and object bounding boxes. The dataset covers a wide range of object categories, including people, animals, vehicles, household items, and more.

In addition to object annotations, the COCO dataset also includes over 2.5 million object instances with segmentation masks, making it a useful resource for researchers and practitioners working on semantic segmentation tasks.

The COCO dataset has been widely used in computer vision research and has been used as the benchmark for some object detection and segmentation challenges, including the COCO Object Detection and COCO Panoptic Segmentation challenges. The dataset has also been used in a variety of applications, such as autonomous driving, robotics, and image retrieval.

IV. CONCLUSION

Real-time object detection is an exciting and rapidly evolving field in computer vision and machine learning. The ability to accurately detect and track objects in real-time video streams has many practical applications, such as in surveillance, robotics, and autonomous vehicles.

there are still many challenges and opportunities for future research. For example, improving the accuracy of object detection in complex environments with occlusions and clutter, developing more efficient and robust models for low-power devices, and exploring new applications and use cases for real-time object detection.

As technology continues to advance, we can expect to see even more powerful and sophisticated real-time object detection systems in the future.

REFERENCES

[1]. Mansoor, A., Porras, A. R., and Linguraru, M. G. (2019). “Region proposal networks with contextual selective attention for real-time organ detection,” in 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), (Venice: IEEE), 1193–1196. doi: 10.1109/ISBI.2019.8759480

[2]. Chen, Y., Li, W., Sakaridis, C., Dai, D., and Van Gool, L. (2018). "Domain adaptive faster R-CNN for object detection in the wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake, UT, 3339–3348. : 10.1109/CVPR.2018.00352

[3]. P. Devaki, S. Shivavarsha, G . Bala Kowsalya, M Manjupavithraa, E.A. Vima (2019).”Real-Time Object Detection using Deep Learning and Open CV” International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-12S, October 2019

[4]. R. Girshick, “Fast R-CNN,” in IEEE International Conference on Computer Vision (ICCV), 2015

[5]. J. Redmon, S. Divvala, R.Girshick, A. Farhadi, You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on computer vision and pattern recognition (pp. 779-788)(2016). https://doi.org/10.1109/cvpr.2016.91.

[6]. J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger. In Proceedings of the IEEE Conference on computer vision and pattern recognition(pp.72637271)(2017). https://doi.org/10.1109/cvpr.2017.690.

[7]. Y.M. Wei, J.C. Quan, Y.Q.Y. Hou, Aerial image location of the unmanned aerial vehicle based on YOLO V2[J]. Laser & Optoelectronics Progress, 54(11): 111002(2017). DOI:https://doi.org/10.3788/LOP54.111002.

[8]. Ashwani Kumar1 , Zuopeng Justin Zhang2 and Hongbo Lyu3. Kumar et al. EURASIP Journal on Wireless Communications and Networking (2020) 2020:204 https://doi.org/10.1186/s13638-020-01826-x

[9]. J. Redmon, A. Angelova, Real-time grasp detection using convolutional neural networks. In 2015 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1316-1322). IEEE(2015, May).

[10]. J. Redmon, S.Divvala, R.Girshick, A. Farhadi, You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on computer vision and pattern recognition (pp. 779-788)(2016).

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, Issue 5, April 2023

[11]. Y. Zhong, Y. Yang, X. Zhu, E. Dutkiewicz, Z. Zhou, T. Jiang, Device-free sensing for personnel detection in a foliage environment. IEEE Geoscience and Remote Sensing Letters **14**(6), 921–925 (2017). https://doi.org/10.1109/LGRS.2017.2687938

[12]. Mitchell, T. M. (1997). Machine learning. McGraw Hill Series in Computer Science. Maidenhead: McGraw-Hill.