IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125270

Real-Time Object Detection and Tracking Using Deep Learning Techniques

Arjun Jadhav[1], Ganesh Dakle[2], Aditya Kundhe[3], Parag Vispute[4]

Students, Computer Engineering, D.Y Patil Institute of Engineering and Technology, Ambi, Pune, India[1-4]

Abstract: The operation of construction vehicles in construction and evacuation sites presents unique challenges due to the different driving conditions and surrounding environment compared to traditional transportation vehicles. Implementing autonomous driving for construction vehicles requires addressing these challenges, even though the learning approach is similar to that of cars. This thesis aims to identify suitable and highly efficient Convolutional Neural Network (CNN) models for real-time object recognition and tracking of construction vehicles, evaluate their classification performance, compare the results, and present the findings.

To achieve these objectives, a literature review and experiments were conducted. The literature review identified suitable object detection models for real-time object recognition and tracking, while experiments were performed to evaluate the performance of the selected models. Based on the literature review, Faster R-CNN model, YOLOv3, and Tiny-YOLOv3 were identified as the most suitable and efficient algorithms for detecting and tracking scaled construction vehicles in real-time. The classification performance of these algorithms was calculated and compared with each other, and the results were presented. The evaluation results indicate that YOLOv3 achieved the highest F1 score and accuracy among the algorithms, followed by Faster R-CNN. Therefore, it is concluded that YOLOv3 is the best algorithm for real-time detection and tracking of scaled construction vehicles. These findings align with the classification performance comparison reported in the literature.

Keywords: Object detection and recognition, Deep Learning, Classification performance

I. INTRODUCTION

Object recognition, a fundamental aspect of computer vision, plays a crucial role in enabling autonomous vehicles to identify and classify objects in real-time. Autonomous vehicles, characterized by their ability to sense and respond to the surrounding environment without human intervention, heavily rely on accurate object detection and recognition to detect obstacles and make informed decisions. Therefore, the selection of appropriate algorithms for real-time object detection is of utmost importance.

Various machine learning and deep learning algorithms, such as Support Vector Machine (SVM), Convolutional Neural Networks (CNNs), Regional Convolutional Neural Networks (R-CNNs), and You Only Look Once (YOLO) model, exist for object detection and recognition. However, for autonomous driving applications, it is crucial to choose an algorithm that can achieve real-time object detection with high accuracy. Since machines cannot instantaneously detect objects in images like humans, the selected algorithms need to be fast, accurate, and capable of real-time object detection. This ensures that the vehicle controllers can solve optimization problems at a frequency of at least one per second.

This thesis is part of a collaborative project between the Project Development Research Laboratory (PDRL) at BTH (Blekinge Institute of Technology) and Volvo CE (Volvo Construction Equipment). The primary objective of the project is to train a model capable of recognizing three types of small-scale vehicles: Hauler, Excavator, and Wheeled Loader. This serves as an initial step towards implementing novel ideas related to machine interaction and intelligent machine navigation systems, including autonomous driving, in a scaled site environment. The scaled site, depicted in Figure 1.4, replicates a small-scale construction/excavation site similar to those found in real-world scenarios (Figure 1.5 and 1.6). Construction vehicles and the construction site environment differ significantly from urban transportation environments, each serving unique purposes. Autonomous construction vehicles are particularly ground breaking as they can address labour shortages and perform tasks with minimal errors over extended periods.

This introduction provides an overview of the importance of object detection in autonomous vehicles and the need for real-time and accurate algorithms. It also highlights the collaborative project between PDRL at BTH and Volvo CE, focusing on training a model to recognize small-scale vehicles in a scaled site environment as a stepping stone towards implementing innovative machine interaction and intelligent navigation systems.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125270

II. PROBLEM STATEMENT

The project aims to address the challenge of real-time object detection, where the goal is to develop an efficient and accurate system that can detect and locate objects within video streamsin real-time. The problem involves overcoming the computational and time constraints associated with processing large amounts of visual data and achieving high detection accuracy simultaneously. The objective is to design and implement a real-time object detection solution that can be deployed in various applications such as video surveillance, autonomous driving, robotics, and augmented reality, providing reliable and timely object detection capabilities.

III. ALGORITHMIC DESIGN

1. CONVOLUTIONAL LAYER

A convolutional neural network (CNN) typically consists of one or more convolutional layers, which can be either pooled or fully connected. The primary purpose of a convolutional layer is to handle computationally intensive tasks. It comprises a set of filters that have the ability to learn. Although the filters are small in size, they span the entire depth of the input. The dimensions of a filter are typically represented as l * w * d, where 'l' represents the height, 'w' represents the width, and 'd' represents the depth of the feature filter, which corresponds to the number of color channels present. During the convolution process, the feature filter slides over the input layer of the neural network, generating a feature map. This layer responsible for the convolution process is called a convolutional layer, and networks consisting of convolutional layers are referred to as convolutional neural networks. As illustrated in Figure 2.4, in the initial stages, the filter searches for specific patterns in the input layer. During the training phase, the filter learns to recognize patterns, which is then utilized to validate the presence of specific patterns during the testing phase. In reality, multiple feature filters exist, each learning to recognize different patterns.

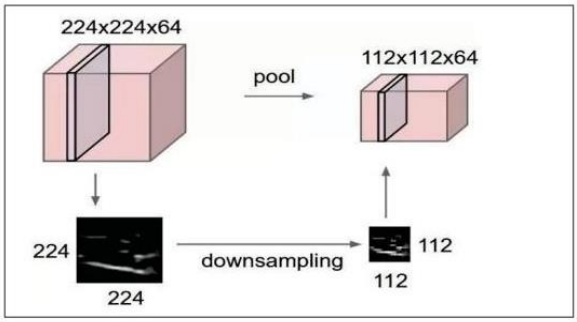

Figure 2.5: Pooling Layer [7]

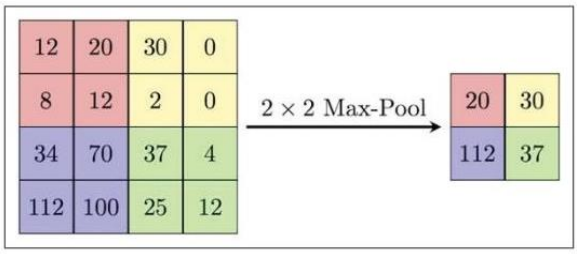

Fig. 1 Example of 2x2 Max-pooling

The state-of-the-art object detector, YOLOv3, is designed to achieve high accuracy while maintaining real-time performance. It is an improvement over the previous version, YOLO. YOLOv3 utilizes a single neural network that predicts the position and class scores of objects in a single iteration. This is achieved by treating the object detection problem as a regression problem, transforming the input images into corresponding class probabilities and positions. YOLOv3 generates multiple S x S grids from the input image, and boundary boxes (B) are predicted, which consist of the height, width, and center coordinates (x and y) of each box. Each box is assigned an object probability (P) and predicts the number of classes (C) within it, along with conditional class probabilities (Pclass) in the S x S grid. The overall prediction of the network is given by S x S x (Bx5 + C), where the digit 5 represents the box coordinates (4) and object probability (1). During the testing phase, the network determines the number of classes present in each grid using equation (2.2). A threshold value, Pmin, is defined at the beginning of the test, and the system detects objects only when Pclass>Pmin. Non-maximal suppression is employed during the post-processing stage to eliminate duplicate detections

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125270

of the same object. The output of YOLOv3 is a tensor with dimensions S x S x (B * 5 + C) [8]. In YOLOv3, bounding boxes have been replaced by 'Anchors' to address the unstable gradient issue that occurred during training. Thus, YOLOv3 predicts outputs with confidence scores by generating a vector of bounding boxes when given an input in the form of an image or a video.

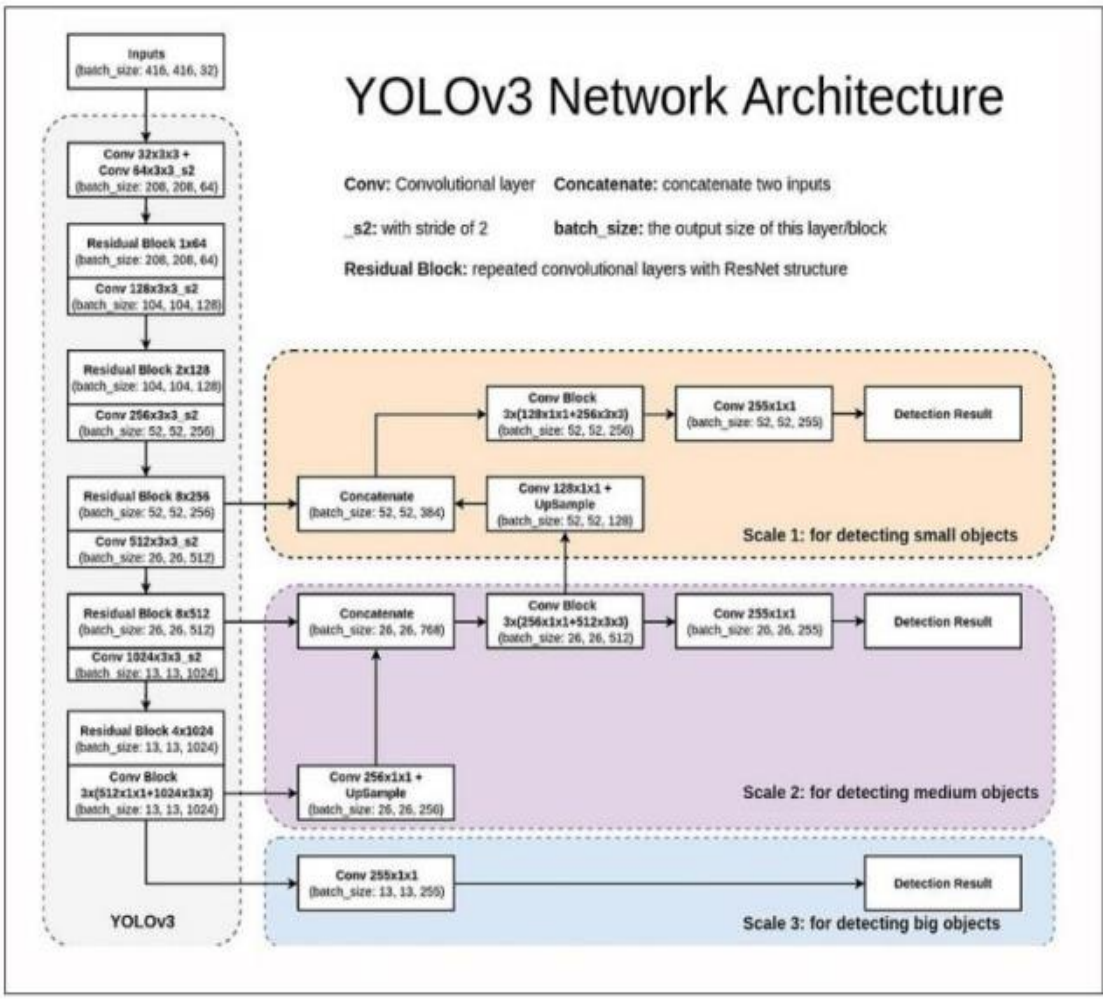

Fig. 2 The Architecture of YOLOv3

2. MATHEMATICAL MODEL

Based on the diverse features of multi-target optical remote sensing image information, the chaotic time series analysis method is employed for classifying and recognizing the multi-target optical remote sensing image information in the mathematical model. The framework of the mathematical model involves expanding and calculating the differential and topological invariants in the original unknown mathematical model in the reconstructed phase space. This establishes a new mathematical model in the reconstructed m-dimensional phase space, enabling prediction, analysis, guidance, and analysis of the original unknown mathematical model [28]. The analysis flow chart, according to chaos theory, is illustrated in the model building analysis process.

Initially, the classification error rate of multi-target optical remote sensing image classification is mapped into a set of probability density functions using a classifier channel mapping function. The m-dimensional vector formed in the reconstructed m-dimensional phase space is represented as follows:

$$Xn = \{x(n), x(n+\tau), ..., x(n+(m-1)\tau)\}$$

Here, n = 1, 2, ..., N, represents a one-dimensional vector Xn of multi-target optical remote sensing images in the m-dimensional phase space of multi-target optical remote sensing image reconstruction mapping. It is expressed as a point

DOI: 10.17148/IJARCCE.2023.125270

in the phase space, and its nearest neighbor point is Xη(n). The variable x denotes the univariate time series of multi-target optical remote sensing images. The distance scale τ, j, and l are constants, and Xj is the vector that measures the distance Rmm between two nodes:

$$Rmm = ||Xj - Xl||$$

How to compare the classification information with the original information to determine the significance of the difference and judge the chaotic probability analysis mapping classification [29]? The chaotic probability of mapping information Qs for multi-target optical remote sensing images is used to analyze the mean ⟨Qs⟩ and compare it with the original Q0. Additionally, the deviation of these Qs values is considered. When the standard deviation σsis fixed and ⟨Qs⟩ is 0, a larger deviation σs indicates more scattered σs values [30]. Some Qs values may be very close to the difference of Q0, indicating that the difference between Qs and Q0 is not significant. Conversely, smaller Qs values indicate a more significant difference between Qs and Q0. Therefore, the difference saliency S is defined to characterize the chaotic probability analysis of multi-target optical remote sensing image mapping information and the difference between the original information:

$$S = |\langle Qs\rangle - Q0| / \sigma s$$

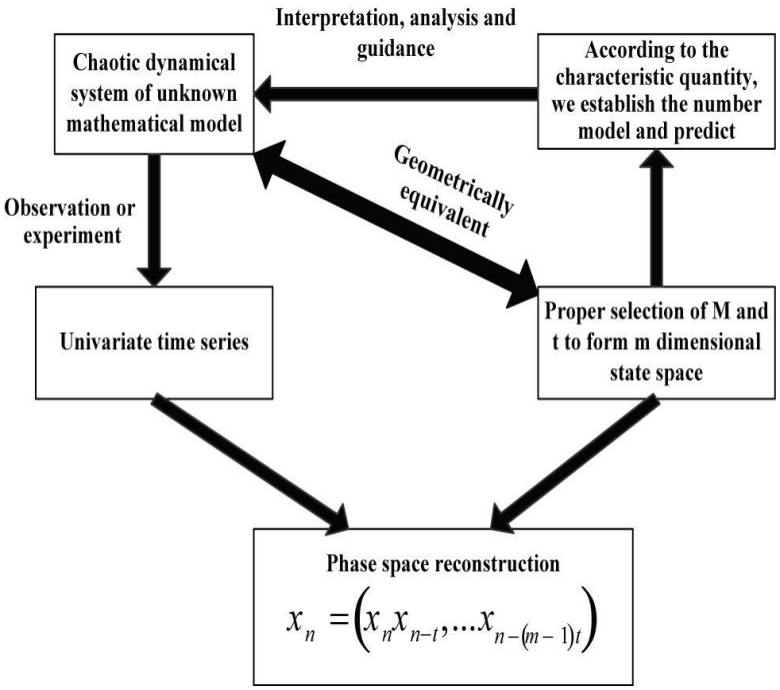

Fig. 3 Mathematical Model

In equation (7), ⟨Qs⟩ represents the mean value of the discriminant statistic obtained from N-batch chaotic probability analysis of multi-target optical remote sensing image mapping information, and σs represents the standard deviation of the discriminant statistic obtained from N-batch chaotic probability analysis of multi-target optical remote sensing image mapping information.

The probability of rejecting the classification based on chaotic probability analysis maps is α = 5%. If the difference between Q0 and ⟨Qs⟩ exceeds a certain critical value Qc, then:

$$p(|Q0 - \langle Qs\rangle| > Qc) \leq 0.05$$

Using the standard normal distribution (σs = 0, ⟨Qs⟩ = 0), the final classification results are obtained as follows:

$$0.025 = \int[-\infty, z2]\ p(Qs)$$
$$dQs = 1 - \int[-\infty, z2]\ p(Qs)\ dQs$$

In equation (11), z2 = -z1 represents the intersection of the rejection region and the confidence interval established by the mapping classification of multi-target optical remote sensing images, and d is a multi-characteristic coefficient.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.125270

IV. CONCLUSION

The real-time object detection project aims to develop a system that can accurately detect and trackobjects in real-time scenarios. Throughout the project, various components and techniques were employed, including data acquisition, pre-processing, object detection, object tracking (if applicable),visualization, and output generation. By leveraging deep learning models and computer vision algorithms, the system can achieve reliable object detection and tracking, providing valuable insightsand applications in different domains.

REFERENCES

[1] R. Philippe, Volvo Excavation Site, 2020. [Online]. Available: www.korestudios.com/portfolio/volvo-construction-equipment/
[2] AHK, Exemplar Construction Site, 2018. [Online]. Available: www.urbantoronto.ca/news/2018/04/torontos-largest-construction-site-well-spadina-front
[3] V. G. Maltarollo, K. M. Honório, and A. B. F. da Silva, "Applications of artificial neural networks in chemical problems," Artificial neural networks-architectures and applications, pp. 203–223, 2013.
[4] TutorialsPoint, Supervised Learning, 2020. [Online]. Available: www.al_network/artificial_neural_network_supervised_learning.html
[5] B. Frank, Deep Learning the Beautiful Mind, 2016. [Online]. Available: www.mindwise- groningen.nl/deep-learning-the-beautiful-mind/
[6] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," inEuropean conference on computer vision. Springer, 2014, pp. 818– 833.
[7] P. Firelord, Pictorial example of max-pooling, 2018. [Online]. Available: https://computersciencewiki.org/index.php/Max-pooling_/_Pooling
[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
[9] CyberAILab, A Closer Look at YOLOv3, 2018. [Online]. Available: https://www.cyberailab.com/home/a-closer-look-at-yolov3

