Proceedings of the Second International Conference on Automation, Computing and Renewable Systems (ICACRS-2023)
IEEE Xplore Part Number: CFP23CB5-ART; ISBN: 979-8-3503-4023-5

2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS) | 979-8-3503-4023-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACRS58579.2023.10404710

Enhancing Surveillance Systems with YOLO Algorithm for Real-Time Object Detection and Tracking

Anish A[1]
UG Student
Department of ECE
Saveetha Engineering College
Chennai,India
anish67657@gmail.com

Sharan R[2]
UG Student
Department of ECE
Saveetha Engineering College
Chennai, India
sharanr2453@gmail.com

Ms.A. Hema Malini[3]
Assistant Professor
Department of ECE
Saveetha Engineering College
Chennai,India
hemamalini@saveetha.ac.in

Ms.T.Archana[4]
Assistant Professor
Department of ECE
Saveetha Engineering College
Chennai,India
archana@saveetha.ac.in

***Abstract*-A Virtually Impaired Person (VIP) is unable to identify objects when they cannot recognize where the object is placed. The researchers are working on it to enhance object detection and help VIP. The challenges faced by researchers are performing detection under low-resolution images, insufficient sensors, portability, and cost. Making a compact device and alerting them is required. By considering the above-mentioned difficulties, an innovative solution is described in this research work. The growth of image processing and deep learning techniques has simplified the complexity of processing data and provided accurate results within a limited time period. The suggested technique presented is a deep learning algorithm called the YOLO algorithm, which is combined with the web to predict objects accurately. For this approach, a dataset with a total of 500 images was chosen and trained. The proposed classifier result is satisfactory, and it achieved an overall accuracy of 94%. Furthermore, this proposed technique provides enough output in comparison with several other machine learning and image processing algorithms.**

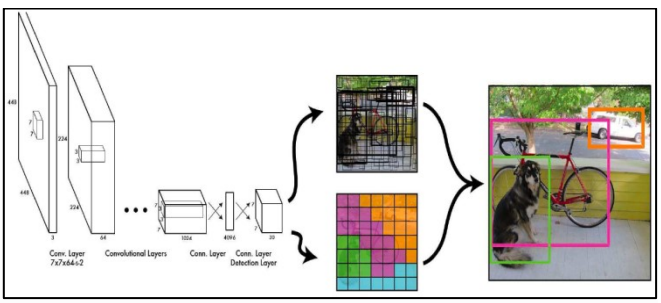
Keywords: Visually Impaired Person (VIP), YOLO Algorithm, Object Detection, Image Processing, Deep Learning.

I. INTRODUCTION

Recently, computational vision and its functionality have been used everywhere, especially in the automobile industry, robots, the healthcare industry, and surveillance systems. Deep learning has garnered significant attention for its remarkable performance in areas like natural language processing, image classification, and object detection. Market projections indicate substantial growth in the coming years, with easy access to powerful Graphics Processing Units (GPUs) and extensive datasets cited as key drivers [1]. Notably, both of these prerequisites have become readily available in recent times [1].

Object detection relies heavily on image classification and recognition, with numerous datasets at our disposal. Microsoft COCO stands out as a widely utilized benchmark for object detection, providing a vast dataset for image classification [2]. In this research study, the author performed a comparative analysis of three prominent object detection algorithms: SSD, Faster-RCNN, and YOLO. SSD enhances detection capabilities by adding multiple feature layers to the network's end, facilitating improved object recognition [3]. Faster R-CNN offers a unified, faster, and more accurate approach to object identification through the use of convolutional neural networks. On the other hand, YOLO, designed by Joseph Redmon, presents an end-to-end network for object detection [3].

This study utilizes Microsoft COCO dataset as a common benchmark and employ consistent evaluation metrics across all three algorithms. This approach enables a fair comparison of the performance of these algorithms, each employing distinct architectural approaches. The results obtained from this comparative analysis offer valuable insights into the unique strengths of each algorithm, allowing us to differentiate their characteristics and determine the most effective object recognition method for specific scenarios.



Architecture of the YOLO

II. RELATED WORKS

Object detection is a crucial research area that leverages the availability of powerful learning tools to explore deeper features. Its goal is to consolidate information on diverse object recognition techniques and classifiers employed by various researchers, facilitating a comparative analysis and practical insights for object detection applications. This work is underpinned by a comprehensive literature review.

Ross Girshick's contributions introduced the Fast R-CNN model, a novel approach to object identification that utilizes Convolutional Neural Networks (CNN). What sets Fast R-CNN apart is its window extraction algorithm, which is different from the traditional sliding window procedure used in the R-CNN model. Fast R-CNN merges individual training for deep convolution networks for feature extraction and Support Vector Machines (SVM) for classification, efficiently combining feature extraction and classification in a unified framework. Remarkably, Fast R-CNN achieves a training time that is nine times faster than R-CNN. Additionally, the Faster R-CNN model integrates the components of proposal isolation and Fast R-CNN into a network template known as the region proposal network (RPN). This achieves accuracy equivalent to that of Fast R-CNN. Collectively, these methods represent a deep learning-based object recognition approach capable of operating at 5–7 frames per second (fps) [4]. This research has provided essential insights into R-CNN, Fast R-CNN, and Faster R-CNN, serving as an inspiration for our model's training.

Kim et al.'s notable work employs CNN in combination with background subtraction to construct a system for detecting and recognizing movable objects recorded on CCTV cameras. The approach hinges on applying the background subtraction classifier to every frame, which informed a similar architecture utilized in our project [5].

Joseph Redmon and his team introduced YOLO, a convolutional neural network architecture that offers a one-stop solution for frame position prediction and the categorization of multiple candidates. YOLO addresses object detection as a regression problem, streamlining the process from image input to category and position output [6]. The methods used in our YOLO architecture for bounding box recognition and feature extraction were inspired by the techniques outlined in this study.

Tanvir Ahmed and their team introduced an innovative YOLO v1 network model, which involved optimizing the loss function, introducing a new inception model structure, and incorporating specialized pooling pyramid layers. This led to improved performance on the PASCAL VOC dataset [7]. Our project utilized this research as a foundation for applying the YOLO model and its training techniques.

Wei Liu and colleagues presented the Single Shot MultiBox Detector (SSD), a novel approach for image object detection. SSD simplified the process by combining object proposal generation and pixel resampling into a single step [8]. Our project adopted training and model analysis methods inspired by their work.

Another research paper introduced a variation of SSD called Tiny SSD. It's a compact single-shot detection deep convolutional neural network designed for real-time embedded object identification. Tiny SSD includes enhanced layers and has a small size of 2.3 MB, making it suitable for embedded applications [9]. In our study, we used a similar SSD model for comparative analysis.

III. PROBLEM STATEMENT

Object detection technology has a wide range of applications, such as autonomous driving, detecting objects from the air, recognizing text, surveillance, assisting in rescue operations, robotics, facial recognition, identifying pedestrians, creating visual search engines, computing objects of interest, and recognizing brands. However, there are several significant challenges to address for its effective implementation.

Variation in Object Occupancy: Objects in images can vary significantly in terms of their size, ranging from taking up a majority of the pixels (70% to 80%) to occupying very few pixels (≤10%).

Multiple Object Sizes: Images often contain objects of various sizes, and detecting objects of different scales can be a complex task.

Labelled Data Availability: Training object detection models requires large volumes of labeled data, and obtaining such data can be resource-intensive and time-consuming.

Object detection using machine learning and deep learning algorithms faces several common challenges. Some of the frequently encountered issues include:

1) Multi-Scale Training: Most object recognition systems are designed and trained for specific input resolutions, which causes them to underperform when presented with inputs of varying scales or resolutions.

2) Foreground-Background Class Imbalance: The presence of an imbalance or disproportion among instances of various object categories can significantly affect the functionality of the suggested approach. Some categories may be overrepresented or underrepresented in the training data.

3) Detection of Smaller Objects: Algorithms trained on larger objects tend to perform well with such objects but often exhibit poor performance when it comes to detecting relatively smaller-sized objects.

To improve the robustness and applicability of object detection algorithms across various domains and applications, it is crucial to address these challenges. Researchers and engineers are actively working on developing innovative solutions to tackle these issues and enhance the performance of object detection systems.

IV. PROPOSED SYSTEM

The block diagram represented in figure 1 the flow of proposed study.

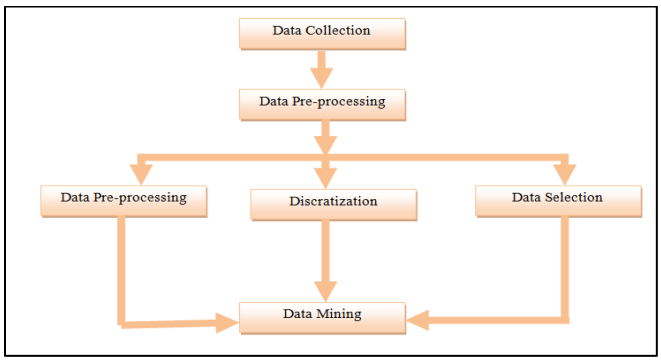


Fig.1.Proposed system

It is an innovative technique to perform object classification by reframing the problem as a regression task instead of a classification problem. In the YOLO architecture, CNN is employed to recognize all bounding boxes and the class probabilities of each object within an image. This unique method allows YOLO to identify objects and their precise positions in a single pass, leading to its name, "You Only Look Once."

The CNN plays a crucial role in feature extraction from visual input, as it efficiently propagates low-level features from initial convolutional layers to deeper layers in a deep CNN. There are several circumstances involved in accurately identifying multiple objects and determining their exact positions within a single visual input. YOLO leverages two key CNN features: parameter sharing and multiple filters, to effectively address these object classification difficulties.

In the recognition process, an image or frame is divided into a grid comprising $S \times S$ cells. Each of these grid cells is responsible for recognizing B-bounding boxes, including their positions, dimensions, the probability of an object's presence within the cell, and conditional class probabilities. The core principle behind object recognition within a grid cell is that the object's centre should be located within that particular cell. Subsequently, the grid cell is responsible for identifying the object using an appropriate bounding box. To be more specific, YOLO detects a set of parameters for a single bounding box in each grid cell. The first five parameters are specific to that particular bounding box, while the remaining parameters are shared among all bounding boxes within the grid, regardless of the number of boxes.

YOLO is an approach based on convolutional neural networks (CNN), and its performance is evaluated on the PASCAL VOC detection dataset.

The YOLO model consists of 24 convolutional layers followed by 2 fully connected layers, highlighting its efficacy in object detection tasks.

V. RESULT AND DISCUSSION

The YOLO structure is presented, which is considered the backbone network for feature extraction. Detection of objects in a room using live-streaming video frames is done with YOLO. The concept of a region proposal network (RPN) suggests the identification of potentially significant regions within an image or frame by generating a redundant collection of overlapping bounding boxes. These proposed regions serve as candidate areas for further examination. Subsequently, a trained model attempts to classify the object type within each bounding box.

In traditional object detectors, the classifiers often analyze the same portion of an image multiple times, which can be computationally intensive. YOLO, however, distinguishes itself by its ability to examine a specific portion of an object just once, unlike available networks. As an object detection technique, YOLO offers a faster processing speed while maintaining comparable precision in object detection tasks.

The implementation of YOLO in object detection is often complemented by tools like OpenCV for displaying the results. By applying Yolo v3, objects within an image or frame can be effectively detected, and their accuracy can be quantified through the use of a confusion matrix. This allows for the evaluation and assessment of the object detection model's performance.

Dataset Collection: In this, raw data is chosen as input. It is unprocessed data.

Data pre-processing: Many sections of the data may be unnecessary or incomplete. Data cleansing is used to manage this aspect. It entails dealing with missing data, noisy data, and so on. This is the most significant module since, without it, the outcome prediction could lead to an inaccurate stage.

Feature Extraction: This stage helps to extract the required features from several features to perform classification by applying a dataset. The original set of features can describe the majority of the data in the new reduced feature set. From a mixture of the original set, a summarized model of the original features is generated. To detect fraudulent behaviour, YOLO extracts both individual and frequency-based behaviour in this module.

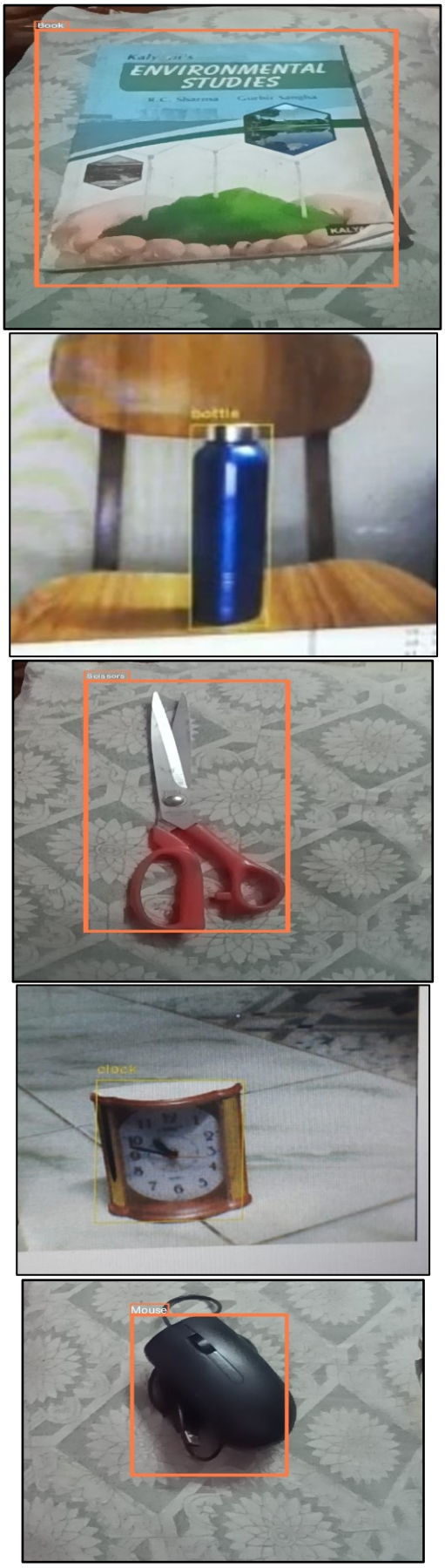


Fig.2.Objects identified in a room

The experimental analysis depicts and names recognized objects such as bottles, clocks, scissors, books, and mouse with various probabilities and repeatedly tracks them with a bounding box of green colour.

VI. CONCLUSION

The proposed system has been developed with the aim of improving object detection and identification in videos, and the findings presented in this thesis demonstrate its enhanced capability in achieving this goal. The research involved a series of experiments that analyzed several approaches to object detection and identification. The study not only establishes a theoretical foundation but also provides practical insights into the efficiency of these methods. As a result, the proposed system offers a comprehensive and well-rounded exploration of object detection techniques, which can be effectively applied in various real-world scenarios.

REFERENCE

1. A. Tiwari, A. Kumar, and G. M. Saraswat, "Feature extraction for object recognition and image classification," *International Journal of Engineering Research & Technology (IJERT)*, vol. 2, pp. 2278–0181, 2013.
2. J. Yan, Z. Lei, L. Wen, and S. Z. Li, "The fastest deformable part model for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2497–2504, New York, NY, USA, 2014.
3. T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik, "Fast, accurate detection of 100,000 object classes on a single machine," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1814–1821, New York, NY, USA, 2013.
4. P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
5. C.-J. Du, H.-J. He, and D.-W. Sun, "Object classification methods," in *Computer Vision Technology for Food Quality Evaluation*, pp. 87–110, Elsevier, Berlin, Germany, 2016.
6. K. W. Eric, Li Yueping, N. Zhe, Y Juntao, L. Zuodong, and Z. Xun, "Deep fusion feature based object detection method for high resolution optical remote sensing images," *Applied Science*, vol. 34, 2019.
7. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
8. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedigs of the International Conference on Computer Vision & Pattern Recognition (CVPR'05)*, pp. 886–893, Berlin, Germany, 2005.
9. 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*, pp. 779–788, Las Vegan, NV, USA, 2016.
10. Y. Zheng, C. Zhu, K. Luu, C. Bhagavatula, T. H. N. Le, and M. Savvides, "Towards a deep learning framework for unconstrained face detection," in *Proceedings of the 2016 IEEE 8th International Conference on BiometricsTheory, Applications and Systems (BTAS)*, pp. 1–8, IEEE, New York, NY, USA, 2016.
11. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, New York, NY, USA, 2014.
12. R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, Berlin, Germany, 2015.
13. W. Liu, D. Anguelov, D. Erhan et al., "Single shot multibox detector," *European Conference on Computer Vision*, vol. 45, pp. 21–37, 2016.
14. T.-Yi Lin, P. Dollár, R. B. Girshick et al., "Feature pyramid networks for object detection," *IEEE CVPR*, vol. 43, pp. 936–944, 2017.
15. W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," *Computer Vision-ECCV 2016*, vol. 43, pp. 21–37, 2016.