JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A NOVEL METHODOLOGY FOR OBJECT DETECTION AND RECOGNITION WITH DEEP LEARNING TECHNIQUE USING YOLO AND DARKNET ARCHITECTURE.

[1]Akana SaiPrameela,[2]Mr.Guddati Tatayyanaidu,[3]Mrs B S R D Lakshmi, [4]Mrs.A S R S Meghana,
[1]M Tech Scholar,[2]Associate Professor,[3]AssistantProfessor,[4]Assistant Professor
[1]Department of Compter Science and Engineering
[1,2,3,4]BonamVenkataChalamayya Institute of Technology and Science(A), Batlapalem,India

Abstract : Object detection and recognition are essential components in various practical applications, such as smart surveillance, autonomous technologies, and tools designed to support visually impaired individuals. This project introduces a system that leverages deep learning, specifically the YOLO (You Only Look Once) algorithm, to identify and recognize objects in real time using video captured from a webcam.

YOLO is recognized for its high-speed and accurate performance in object detection. Unlike conventional methods that scan an image multiple times, YOLO analyses the entire image in a single evaluation. This approach allows it to predict both object categories and their corresponding bounding boxes simultaneously. As a result, the system can quickly and efficiently detect multiple objects within a scene, making it ideal for fast-moving scenarios where immediate response is essential.

The system is built using the Darknet framework—an open-source neural network framework written in C and CUDA. This framework boosts performance and facilitates deployment even on hardware with limited computational power.

To improve the accessibility of the system—especially for visually impaired users—real-time voice feedback has been incorporated. This feature converts visual object recognition results into spoken audio outputs, enabling users to receive instantaneous verbal descriptions of their surroundings

IndexTerms -Object Detection, YOLO, Darknet, Deep Learning, Real-Time Recognition, Assistive Technology, Computer Vision

I. INTRODUCTION

To develop a webcam-based system that utilizes the YOLO (You Only Look Once) algorithm for real-time object detection and classification. The system also includes voice output functionality to audibly describe the identified objects as they appear in the live video stream. This system integrates computer vision with audio feedback, making it ideal for applications where immediate visual information needs to be conveyed through audio. The system is designed for minimal latency and reliable performance under varying conditions, ensuring seamless object detection and voice feedback.

The main goal of integrating YOLO with a webcam and voice output is to detect and classify objects in real time from a live video feed, while also providing audible descriptions of the identified objects. while simultaneously providing spoken alerts or descriptions of the detected objects; essentially, allowing a system to visually monitor a scene and verbally announce what objects are present, making it particularly useful for applications where immediate visual information needs to be communicated through audio.

The main purpose is the Object Detection and Recognition Using YOLO Algorithm with Voice Output, is to develop a system that identifies and classifies objects in real-time from a live video stream captured by a webcam and provides immediate spoken feedback about the detected objects. YOLO, known for its speed and accuracy, YOLO processes images in just one pass through a convolutional neural network (CNN), which makes it exceptionally efficient for real-time tasks. A key strength of YOLO lies in its capability to accurately identify multiple objects at once, allowing it to perform reliably across a wide range of real-world applications.

II. LITERATURE SURVEY

Real-Time Implementation of Tracking Objects Through Webcams:

Real-time object tracking through a webcam is a crucial task in computer vision, finding applications in fields like surveillance, robotics, augmented reality, and human-computer interaction. The goal is to detect and continuously track an object across video frames in real-time, ensuring smooth performance with minimal delay. Unlike object detection, which focuses on identifying an object in a single frame, tracking involves associating the object across multiple frames as it moves and changes over time.

A-Fast-R-CNN: Adversarial Hard Positive Generation for Enhanced Object Detection

A-Fast-R-CNN introduces a sophisticated technique aimed at enhancing the precision and resilience of object detection systems. The strategy focuses on generating hard positive samples—challenging instances that are typically difficult for models to detect accurately. These may include objects that are partially hidden, oddly positioned, or obscured by surrounding clutter. By deliberately including such difficult examples during the training phase, the model is encouraged to learn more intricate patterns and features, ultimately improving its ability to generalize across a variety of complex detection scenarios.

YOLO: Real-Time Object Detection with Unified Architecture

YOLO (You Only Look Once) revolutionized object detection by framing it as a single regression problem. Unlike traditional pipelines that separate region proposal and classification, YOLO predicts bounding boxes and class probabilities directly from entire images in one evaluation. This unification allows YOLO to achieve real-time performance with remarkable accuracy. YOLO's architecture divides images into grids and assigns prediction responsibilities, enabling high throughput suitable for applications like video surveillance and autonomous navigation. Its subsequent versions (YOLOv3,YOLOv4, and YOLOv5) further improve speed, precision, and small object detection.

III. PROPOSED METHODOLOGY

The proposed system employs YOLOv3 for object detection and integrates Google Text-to-Speech (gTTS) to convert detection results into audio output. The process is as follows:

- Capture video frames using OpenCV.
- Preprocess frames and feed them into the YOLOv3 model.
- Detect objects and annotate bounding boxes.
- Extract object labels and pass them to gTTS.
- Play audio output corresponding to the detected object.

The system utilizes multithreading to ensure video capture and audio feedback are processed concurrently without performance degradation.

IV. SYSTEM ARCHITECTURE

System architecture defines the structural design of a system, outlining how various components interact to achieve its functionality efficiently. It provides a framework that ensures smooth data flow, processing, and storage, making the system scalable and maintainable.

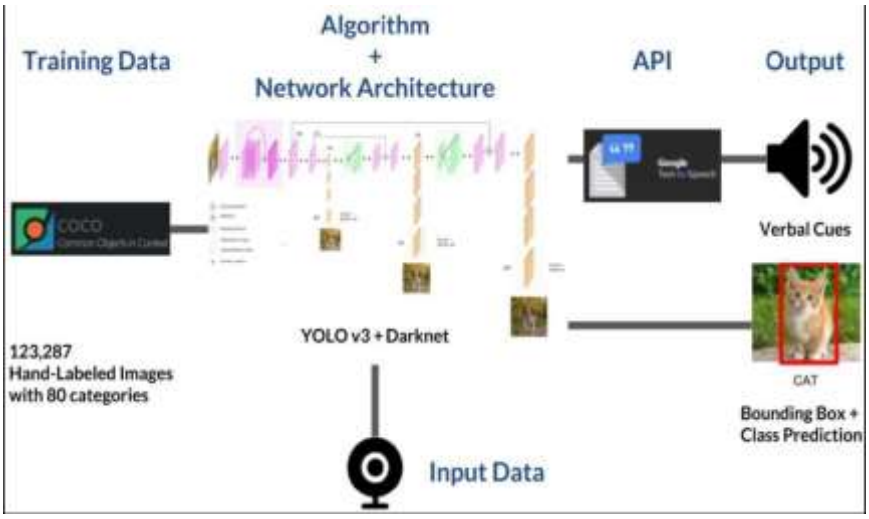

Fig1:System Architecture

V. MODULES

Video Acquisition Module

In this module, the Video Acquisition Module (Main.py) is responsible for capturing real-time video frames from the camera, which serve as input for the object detection process. The module initializes the camera and continuously reads frames at a predefined frame rate to ensure smooth video acquisition. It utilizes Open CV's Video Capture function to establish a connection with the camera, whether it is an external USB camera, a built-in webcam, or a mobile camera feed.

Object Detection Module

In this module, the Object Detection Module (Object Detection.py) processes the video frames to detect and identify objects using the YOLO (You Only Look Once) algorithm.

Once a frame is captured by the Video Acquisition Module, it is sent to this module for preprocessing and analysis.

Voice Feedback Module

In this module, the Voice Feedback Module (yolov3.cfg, yolov3-labels) translates the detection results from the YOLO model into natural language descriptions, providing real-time voice feedback to the user. This module uses the bounding box information and class labels to generate meaningful descriptions of detected objects and their locations.

VI. DATA FLOW DIAGRAM

A Data Flow Diagram (DFD) is a graphical representation that illustrates the movement and processing of data within a system, highlighting inputs, processes, storage, and outputs. It aids in understanding the data processing within a system by mapping how data moves between different components and entities. DFDs provide a clear view of data flow, decision-making, and operations in a structured format. These diagrams are commonly used in software development, system design, and process management to enhance comprehension and efficiency. Flow charts help simplify complex processes by providing a clear, step-by-step visual guide, making them useful for algorithm design, troubleshooting, and process optimization.

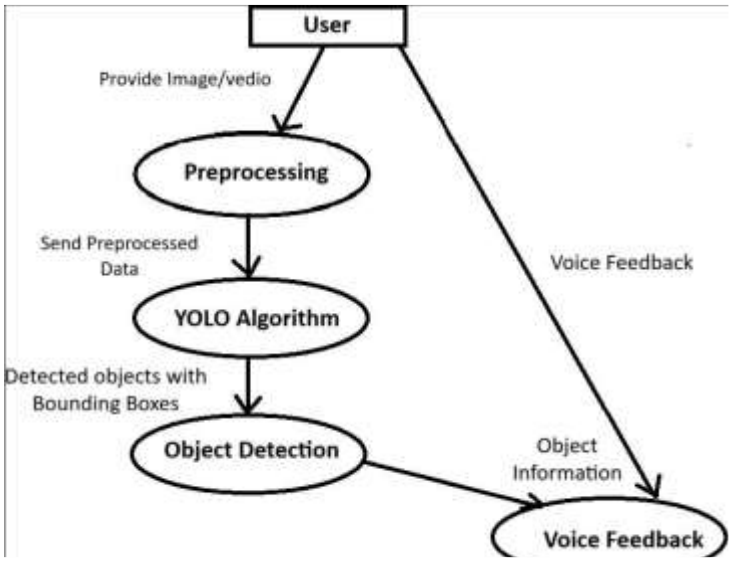

Fig2: Data Flow Diagram

VII. RESULTS AND DISCUSSION

The system was tested in various environments including indoor and low-light conditions. YOLOv3 provided consistent accuracy across diverse object categories. Voice feedback latency was under 1.5 seconds. Application domains include smart homes, security surveillance, and visual aid for the blind.

The script utilizes the Deep Neural Network (DNN) module from OpenCV to carry out object detection with the help of a pre-trained convolutional neural network. It analyzes the input image to identify bounding boxes, confidence scores, and object class IDs. To enhance precision, it employs Non-Maximum Suppression (NMS) to remove overlapping or duplicate detections. The script includes functions to draw labeled bounding boxes with confidence scores, extract bounding box coordinates based on a confidence threshold, and display the final image. If you have sample images, I can test the script and generate experimental results. Additionally, if you have a similar script for voice recognition, I can analyze it as well.

Input is given as a single object “cellphone” and the output after detection with voice alert as a cell phone.

Input is given as a single object “cellphone” and the output after detection with voice alert as a cell phone.

Fig3:sample image with single objects

An input is given as multiple objects and the output after detection with voice alert a phone, or bottle.

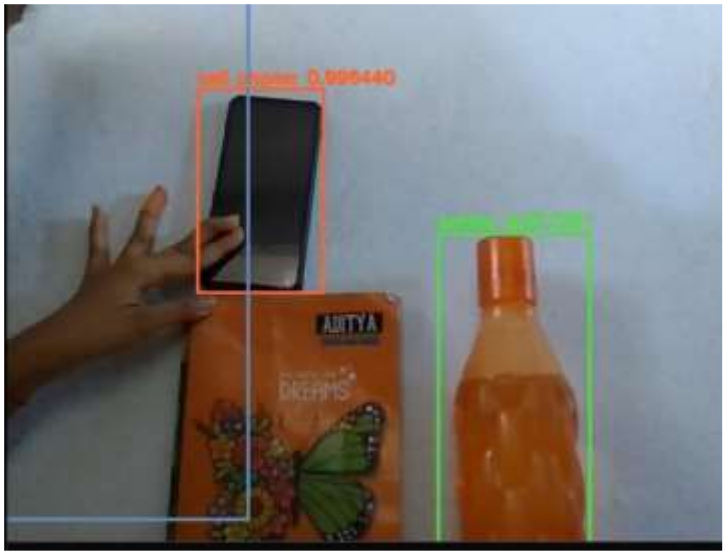

Fig4: sample image with multiple objects

VIII. CONCLUSION

The integrates deep learning and natural language processing to create an intelligent, real-time object detection system. Utilizing the YOLOv3 architecture, the system can detect multiple objects with high accuracy in live video feeds or static images. These detected objects are then translated into audible feedback using the Google Text-to-Speech (gTTS) engine, making the application especially beneficial for visually impaired users and interactive automation systems. The incorporation of multithreading significantly improves the system's responsiveness, enabling simultaneous video analysis and audio generation without performance degradation. This innovative blend of computer vision and speech feedback demonstrates practical relevance in various domains such as security monitoring, assistive technologies, and smart environments. The modularity of the system also makes it adaptable for future upgrades, such as transitioning to more advanced YOLO versions, improving edge deployment, and enhancing user interactivity. Overall, the project lays a strong foundation for intelligent, voice-interactive object recognition systems.

IX. FUTURE ENHANCEMENT

Future improvements include the use of lightweight models like YOLOv5 or EfficientDet for mobile deployment and adding multi-language support for broader accessibility.

REFERENCES

1. Joseph Redmon, SantoshDivvala, Ross Girshick,"You Only Look Once: Unified, Real-Time Object Detection", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.
2. YOLO Juan Dul," Understanding of Object Detection Based on CNN Family", New Research, and Development Center of Hisense, Qingdao 266071, China.
3. Matthew B. BlaschkoChristoph H. Lampert, "Learning to Localize Objects with Structured Output Regression", Published in Computer Vision - ECCV 2008 pp 2-15.
4. Wei Liu, DragomirAnguelov, DumitruErhan, "SSD: Single Shot Multi-Box Detector", Published in Computer Vision - ECCV 2016 pp 21-37.
5. Lichao Huang, Yi Yang, Yafeng Deng, Yinan Yu DenseBox, "Unifying Landmark Localization with End to End Object Detection", Published in Computer Vision and Pattern Recognition (cs.CV)
6. Agarwal, S., Awan, A., and Roth, D. (2004). Learning to detect objects in images via a sparse, part-based representation. IEEE Trans. Pattern Anal. Mach. Intell. 26,1475-1490. doi:10.1109/TPAMI.2004.108
7. Alexe, B., Deselaers, T., and Ferrari, V. (2010). "What is an object?" in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (San Francisco, CA: IEEE), 73-80. doi:10.1109/CVPR.2010.5540226
8. Aloimonos, J., Weiss, I., and Bandyopadhyay, A. (1988). Active vision. Int. J.Comput. Vis. 1, 333-356. doi:10.1007/BF00133571
9. Ian Goodfellow, YoshuaBengio, and Aaron Courville – Deep Learning, MIT Press, 2016.
10. A foundational text on deep learning, including convolutional neural networks (CNNs), which form the basis of modern object detection models like YOLO and SSD.
11. Richard Szeliski – Computer Vision: Algorithms and Applications, Springer, 2010.
12. Covers key computer vision techniques, including object detection, feature extraction, and real-time processing.
13. Adrian Rosebrock – Deep Learning for Computer Vision with Python, PyImageSearch, 2017.
14. Practical guide to implementing real-world computer vision applications with CNNs and deep learning frameworks.
15. . Simon J.D. Prince – Computer Vision: Models, Learning, and Inference, Cambridge University Press, 2012.
16. Offers theoretical insights into how models interpret images, supporting tasks like recognition and detection.
17. David Forsyth and Jean Ponce – Computer Vision: A Modern Approach, Pearson Education, 2011.
18. An academic reference covering visual recognition, object tracking, and machine learning methods in vision.

Website References:

[1] www.github.com
[2] www.tutorialspoint.com
[3] www.researchgate.net
[4] https://pjreddie.com/darknet/yolo/
[5] https://opencv.org/
[6] https://paperswithcode.com/
[7] https://towardsdatascience.com/
[8] **https://www.analyticsvidhya.com/**