Visually Empowered: Real-Time Object Recognition and Auditory Description System

Paruchuru Kundhana Harshitha
Indiana University, Bloomington
Bloomington, Indiana, USA
kparuch@iu.edu

Abstract

This research project focuses on leveraging computer vision and natural language processing techniques to enhance accessibility and improve the quality of life for visually impaired individuals. The data collection phase involves working with the COCO dataset, a benchmark dataset for object detection tasks, and experimenting with various versions of YOLO. Among these models, YOLOv8m demonstrated superior performance by detecting a higher number of objects with higher class confidence. Subsequently, object labels are processed using NLP translation to generate meaningful English sentences. Leveraging the langdetect library for language detection, the English sentences are then translated into the desired language using Google Translate multilingual support. Finally, the translated sentences are read aloud using the Google Text-to-Speech module. This comprehensive integration of cutting-edge technologies aims to provide visually impaired individuals with independent navigation capabilities and facilitate their engagement with the environment. The project's goal is to contribute positively to the social cause of aiding visually impaired individuals in accessing and comprehending visual information effectively.

1. Introduction

Visually impaired individuals face significant challenges in their daily lives, as sight is often considered the most crucial sense organ, providing approximately 80% of our perception [1]. According to the World Health Organization, around 2.2 billion people worldwide experience vision impairments either faraway or close, out of them 49.1 million classified as visually impaired. However, population growth contributes to an increasing number of affected individuals, requires enhanced vision impairment prevention efforts globally. Visually impaired individuals typically rely on stick or assistance from others. Vision impairment also significantly impacts adults, leading to social isolation and increased dependence on care services. This project aims to assist visually impaired individuals in recognizing their surroundings, leveraging deep learning techniques like You Only Look Once (YOLO). These deep learning systems offer high accuracy rates at a reasonable cost, making them effective solutions for object detection and recognition tasks in accessibility technologies.

In recent years, deep learning has gained popularity as an effective technique for object identification, achieving high accuracy rates at a reduced cost [2,3]. Single Shot Detector (SSD) [4] and You Only Look Once (YOLO) [5] have been widely adopted to address detection and recognition challenges.

This paper introduced an end-to-end system to assist visually impaired individuals. It employs the YOLOv8m model for real-time object detection, uses NLP translation to convert the object labels into meaningful English sentence and integrates speech generation tool such as gTTS (Google Text-to-Speech) to convert detected images into spoken text. The contributions of this work include the development of a practical assistive system, leveraging state-of-the-art deep learning models and speech synthesis technology to enhance accessibility and independence for visually impaired individuals.

2. Literature Survey

Atikur Rahman and Sheikh Sadi [6] introduced an IoT-enabled Automated Object Recognition system utilizing the SSD Model, SIFT, and the MS COCO dataset in 2021. Balachandar, Santhosh et al. [7] devised a technique incorporating a multi-view object tracking (MVOT) system to handle multiple cameras for video monitoring and capture, creating a robust and precise framework based on video data. Mansi Mabendru and Sanjay Kumar Dubey [8] developed a system employing YOLO and YOLOv3 algorithms separately, evaluating their accuracy and performance. The SSD MobileNet model is integrated into the YOLO TensorFlow framework, while the Darknet model is used in YOLOv3. They also utilized the gTTS Python library for transforming text into audio for audio feedback. Kanchan Patil et al. [9] proposed a wearable device featuring a virtual assistant system comprising five

components, which can be navigated using hardware buttons and user-provided voice commands. Natural language processing techniques are essential for generating voice-based image captions, as demonstrated by Mohana Priya et al. [10].

3. Methodology

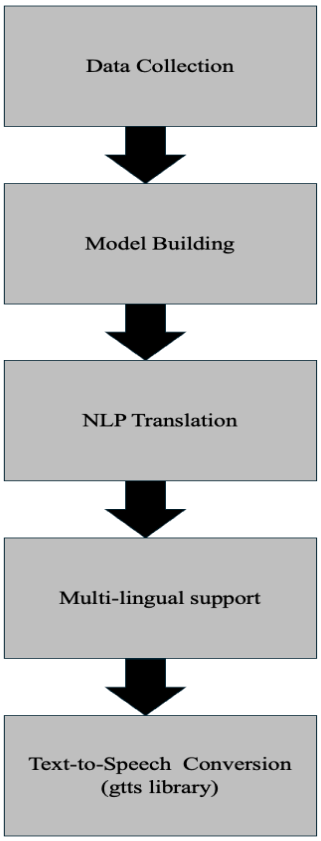


Fig 1. This is the proposed methodology.

3.1 Data Collection

The COCO (Common Objects in Context) dataset is a well-known benchmark in computer vision research, comprising a wide array of images annotated with labels for 80 common object categories. These categories encompass a diverse range of objects commonly found in daily life scenes. The dataset's richness and diversity make it an invaluable resource for training and evaluating models in tasks such as object detection, segmentation, and captioning within the field of computer vision. Fig 2. has objects from different categories.

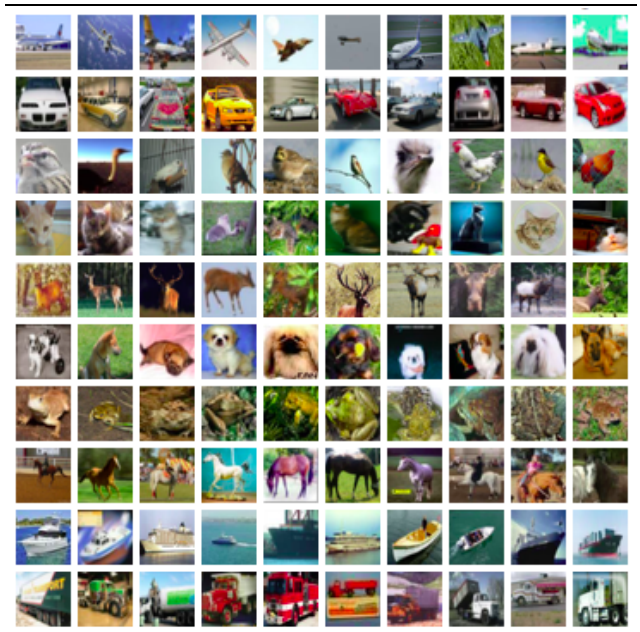


Fig 2. Source: https://cocodataset.org/.

3.2 Model Building

During the model building phase of this research, several versions of the YOLO object detection models were utilized, including YOLOv5nu, YOLOv8m, and YOLOv8n. Each variant of the YOLO model comes with its unique architecture and characteristics, contributing to its performance in object detection tasks.

YOLOv5nu is an updated nano model and member of the YOLOv5 series, known for its real-time object detection capabilities with a focus on speed and accuracy. This model prioritizes efficiency without compromising on accuracy, making it suitable for deployment in resource-constrained environments.

On the other hand, YOLOv8m is the micro model and emerged as the standout performer during our evaluations. This variant of the YOLO model demonstrated superior performance, particularly in detecting a wide range of objects with high precision and confidence. Its robustness and effectiveness in object recognition tasks were notable, showcasing its potential for practical applications.

YOLOv8n is the nano model and another member of the YOLO model series, was also considered and evaluated. However, in our comparative assessments, YOLOv8m consistently outperformed YOLOv8n and other variants, establishing itself as the model of choice for our research project.

The selection of these pre-trained YOLO models was based on their established reputation for accuracy, speed, and versatility in object detection applications. Our thorough experimentation and evaluation process validated YOLOv8m as the most promising model, setting the stage for its potential deployment in real-world object recognition systems.

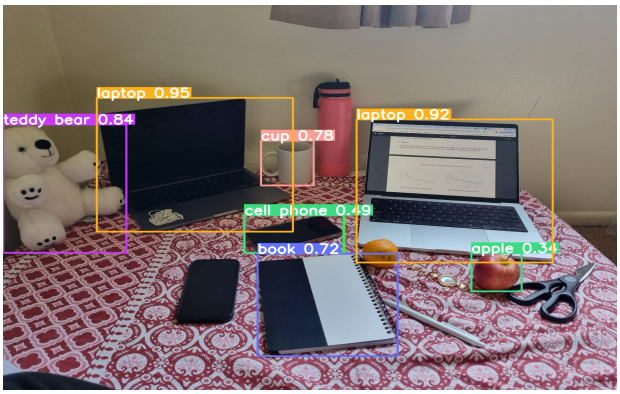


Fig 3. Yolov5nu model object detection

Fig 4. Yolov8n model object detection

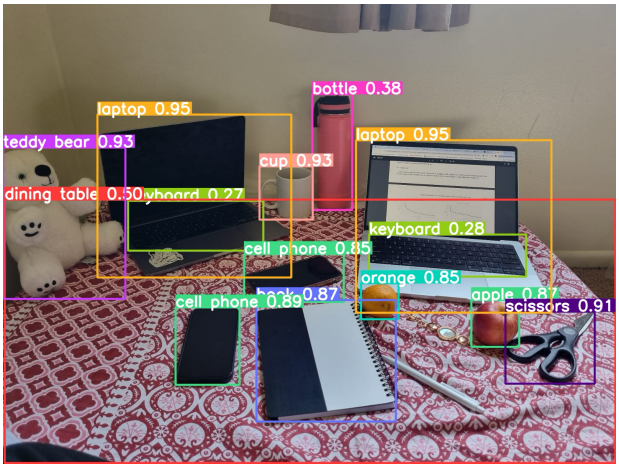


Fig 5. Yolov8m model object detection

Yolov8m outperformed by detecting a greater number of classes with higher class confidence. Yolov8m is only considered in further steps.

3.3 NLP Translation

NLP translation plays a crucial role in converting raw object labels, typically in the form of class identifiers or codes, into meaningful English sentences. This process involves mapping each object label to its corresponding English description or name, providing context and comprehension for the detected objects.

These are the objects identified by Yolov8m model.

```
63 laptop
63 laptop
77 teddy bear
41 cup
76 scissors
67 cell phone
73 book
47 apple
67 cell phone
49 orange
60 dining table
39 bottle
66 keyboard
66 keyboard
```

Fig 6. This are the labels generated by Yolov8m

The text generated from the object labels using NLP translation is “There are 2 laptops, 2 cell phones, 2 keyboards and teddy bear, cup, scissors, book, apple, orange, dining table, and bottle”.

3.4 Multi-lingual support

Employed langdetect and googletrans, to perform language detection and translation tasks. It detects the language of a given text and translates it into another language, for instance, from an english language to Spanish ('es' represents Spanish as the destination language). The Translator class from googletrans is utilized to interface with Google Translate's API, allowing for seamless translation of text strings. The translated output is then extracted and stored for further processing. In Table 1. Some of the languages among 55 are shown.

Language	Text
English	There are 2 laptops, 2 cell phones, 2 keyboards and teddy bear, cup, scissors, book, apple, orange, dining table, and bottle.
German	Es gibt 2 Laptops, 2 Mobiltelefone, 2 Tastaturen und einen Teddybären, eine Tasse, eine Schere, ein Buch, einen Apfel, eine Orange, einen Esstisch und eine Flasche.
French	Il y a 2 ordinateurs portables, 2 téléphones portables, 2 claviers et un ours en peluche, une tasse, des ciseaux, un livre, une pomme, une orange, une table à manger et une bouteille.
Spanish	Hay 2 computadoras portátiles, 2 teléfonos celulares, 2 teclados y un osito de peluche, una taza, tijeras, un libro, una manzana, una naranja, una mesa de comedor y una botella.
Japanese	ラップトップ 2 台、携帯電話 2 台、キーボード 2 台、テディベア、カップ、はさみ、本、リンゴ、オレンジ、ダイニング テーブル、ボトルがあります。

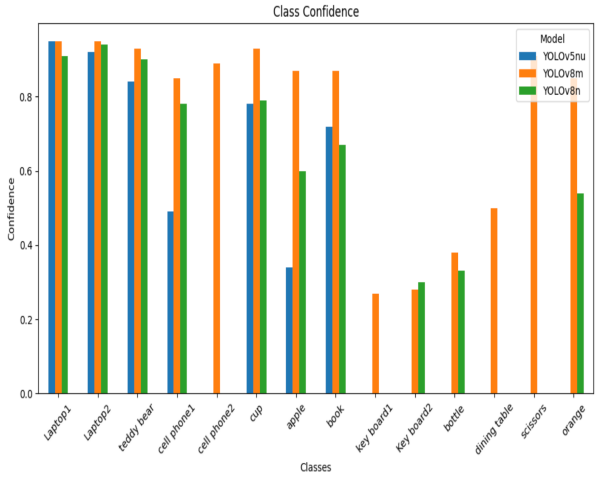
Table 1. Generated in 5 different languages.

3.5 Text-to-Speech Conversion

The gTTS (Google Text-to-Speech) library is a Python package that allows you to convert text into speech using Google's Text-to-Speech API.
Its key features include robust text-to-speech conversion capabilities, supporting multiple languages to cater to diverse linguistic needs. The library provides customization options such as choosing different voices, adjusting speaking rates, and specifying output audio formats like MP3 or WAV, allowing users to tailor the speech output according to their preferences. With a user-friendly API, gTTS simplifies the integration of text-to-speech functionality into Python applications, making it ideal for a wide range of use cases including voice-enabled applications, automated messaging systems, and accessibility solutions.

4. Results

The paper highlights the significance of integrating object detection models with text-to-speech technology for enhancing accessibility for visually impaired individuals. It may emphasize the practical implications of such technology in bridging the gap between visual perception and auditory comprehension, empowering visually impaired individuals to navigate their environment more confidently. YOLOv8m outperformed other models for accurate object recognition and the utility of gTTS for generating meaningful audio descriptions.



5. Future Work

In future, this can be integrated in wearable devices.

References

[1] Center, M.E., 2016. Importance of Eye Care, available at, https://www.medicaleyecenter.com/2016/06/20/importance-eye-care/.URL:https://www.medicaleyecenter.com/2016/06/20/importance-eye-care/. [Online; accessed 6-November-2021].

[2] Guravaiah, K., Rithika, G., Raju, S.S., 2022. Homeid: Home visitors recognition using internet of things and deep learning algorithms, in: 2022 International Conference on Innovative Trends in Information Technology (ICITIIT), IEEE. pp. 1–4.

[3] Guravaiah, K., Velusamy, R.L., 2019. Prototype of home monitoring device using internet of things and river formation dynamics-based multi-hop routing protocol (rfdhm). IEEE Transactions on Consumer Electronics 65, 329–338.

[4] Joshi, R., Tripathi, M., Kumar, A., Gaur, M.S., 2020. Object recognition and classification system for visually impaired, in: 2020 International Conference on Communication and Signal Processing (ICCSP), IEEE. pp. 1568–1572.

[5] Mahendru, M., Dubey, S.K., 2021. Real time object detection with audio feedback using yolo vs. yolo v3, in: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE. pp. 734–740.

[6] Rahman, M.A., Sadi, M.S., 2021. Iot enabled automated object recognition for the visually impaired. Computer Methods and Programs in Biomedicine Update , 100015.

[7] Balachandar, A., Santhosh, E., Suriyakrishnan, A., Vigensh, N., Usharani, S., Bala, P.M., 2021. Deep learning technique based visually impaired people using yolo v3 framework mechanism, in: 2021 3rd International Conference on Signal Processing and Communication (ICPSC), IEEE. pp. 134–138.

[8] Mahendru, M., Dubey, S.K., 2021. Real time object detection with audio feedback using yolo vs. yolo v3, in: 2021 11th International Conference on Cloud Computing,

Data Science & Engineering (Confluence), IEEE. pp. 734–740.
[9] Patil, K., Kharat, A., Chaudhary, P., Bidgar, S., Gavhane, R., 2021. Guidance system for visually impaired people, in: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE. pp. 988–993.
[10] Anu, M., Divya, S., et al., 2021. Building a voice based image caption generator with deep learning, in: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE. pp. 943–948.