Grenze International Journal of Engineering and Technology, June Issue

A Real-Time Assistive Vision System using YOLOv5

Adyaa Khaneja[1], Raghav Gupta[2] and Betty Paulraj[3]
[1-3]Department of CSE, Amity University Uttar Pradesh, Noida, India
Email: adyaa.khaneja@s.amity.edu, raghav.gupta4@s.amity.edu, bpaulraj@amity.edu

***Abstract*— Assistive technologies have become increasingly prominent over the past few years because of their extensive usage. There have been many studies conducted over the last two decades on technological implementation into medicine and the development of assistive technologies geared toward those individuals with disabilities. The implemented technologies may assist individuals with visual impairments with navigation and object detection. The proposed system's software incorporates real-time object recognition, optical character recognition, and speech synthesis capabilities utilizing a single-stage, modified deep learning object detection model designed specifically for real-time object recognition and accurate object identification without latency or privacy concerns. Image input is taken from a live camera feed, the input image is processed, and the object recognition results are provided based on performance data with a low latency rating and privacy level; therefore, the final product is subject to the determined embedded constraints from the tested architecture. The test environment was both controlled (data set) and uncontrolled (i.e., living environment), and data collected during testing was used to evaluate the product's effectiveness and determine the suitability of using low-cost technology to create a vision-based assistive system to field-test products.**

I. INTRODUCTION

Over the course of the last decade, there has been a substantial emphasis on the advancement of assistive technologies. Within healthcare, there are various ways to utilize assistive technologies, all having led to numerous research possibilities. The intent of this project work is to provide practical solutions for visually impaired individuals. In this work, the implementation of Natural Language Processing (NLP) as well as Object Detection will be looked at and implemented into this project work. Many studies have been done within published literature on the different methods such as Machine Learning, Deep Learning, and other variations of existing ML models. There is little research focusing on adding artificial intelligence to different types of assistive technologies. The studies that have been done on state-of-the-art methods have all identified many practical limitations that cause the end user to experience inconvenience when using these systems in everyday life. Therefore, the goal of this project was to develop an affordable, real-time, embedded device for object detection and alerting the end user through speech. With this as our goal, we used YOLOv5 as our object detection model based on our review of existing literature in state-of-the-art methods. Following that, we developed an OCR model and then completed natural language processing before finally producing audio output for the end user.

The proposed system has undergone testing and successful implementation in an embedded environment where time constraints and other real-time considerations or parameters were applied and documented in this paper.

Grenze ID: 01.GIJET.12.2.403_24

The sections of this manuscript will be described in the following paragraphs: Section 1 Introduction/Problem Statement; Section 2 Literature Review (Brief); Section 3 Methodology for the Proposed System; Section 4 Implementation & Results and Discussion; and finally Section 5 Conclusion/References.

II. Literature Survey

In a recent ten years of research there has been much interest in assistive technology for those with visual impairment. Numerous studies and articles have been written regarding the various types of assistive technologies that are available to individuals with visual impairment. In study [1], a report was published which describes how machine learning techniques were able to create vision and navigation systems for individuals with visual impairment. Following this, study [2] describes deep learning techniques being utilized to enable the delivery of real-time object detection for various object detection models and also based on auditory feedback system. Study [3] proposes using deep learning-based convolutional neural networks (CNNs) for real-time object detection; therefore, the object detection model using You Only Look Once (YOLO) has now also been developed [4] and will be discussed in [6][7]. The methodology to use a single shot text detector will also be delineated in studies [8][9]; as well as the mobile cloud architecture will be presented as a variation of CNN for mobile vision applications in studies Title [10][11][12]. The various machine learning models studied in [13][14][15].

Various studies have examined the latest technology and have pointed out a research gap involving the effects of the conversion process from text to audio and how any time delays impact the performance of both processes; regardless of which one is currently in progress, the performance is still unresolved. Therefore, this proposed method involves building upon previous literature to create a system capable of converting text to speech with very little delay, utilizing photographs taken through a device, such as a camera, to create text and/or audio, as well as identifying the visible objects within each photo using YOLOv5. In addition, once the text is identified, the text is converted to speech and produced as an audio file for end users; the system will be tested in a live environment and the results documented.

III. Methodology

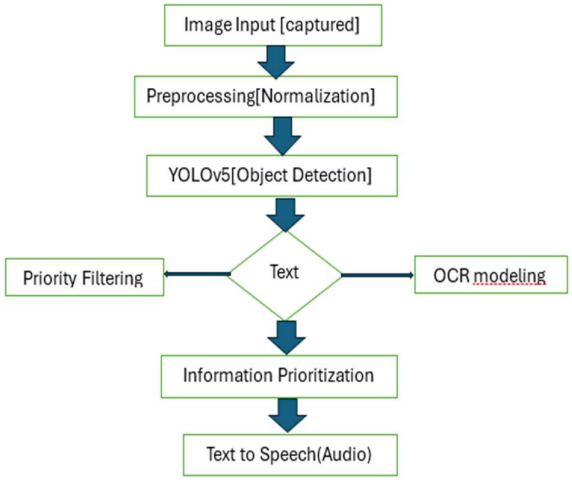

Fig 1. Proposed Methodology Pipeline

Fig 3.1 presents the systematic architecture and processing pipeline for this project, which includes developing a vision assistance system for people with disabilities and adapting it to work in real-time applications.

The system processes images that are received as the input, normalizes them and detects objects using the YOLO V 5 model; once the object is identified and processed through the OCR and filtering priorities, the results will be prioritized at the end stage. The information is then converted from text to speech; as a result we are able to provide ordering abilities back to those who receive them. This will outperform previous methods reported on in the literature by; methodologically identifying gaps created by prior research methods and solutions using this approach.

IV. RESULTS AND DISCUSSION

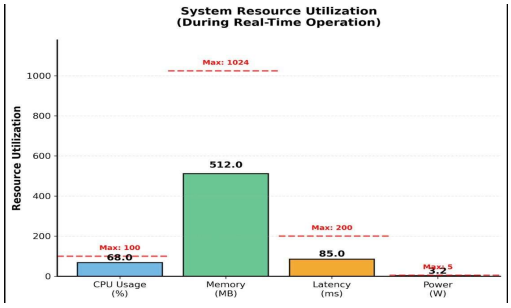

Fig 2. Resource Utilization

We were able to evaluate the success of the system's performance based on its limits: In terms of the flange of a computer (CPU), the maximum capacity was 68%, while the maximum was 100%. For memory, the maximum amount of available memory was 512MB out of an available total of 1024MB of available memory during the test as well. The amount of time it took for the network to send and receive messages averaged 85ms and had a maximum threshold of 200ms; therefore, there are many instances when people who are visually impaired could use the audio feedback quickly. The CPU usage did not exceed 68% during the test of the system, which means there was a 68% capacity during the testing and under the maximum threshold of 100%. During the test of the system, there was also only 512MB of memory used out of a total of 1024MB of available memory. In addition, the average network latency, which is how long it takes to send and receive a message across a network, was 85ms, and had a maximum value of 200ms allowing people who are visually impaired to be able to hear from the assistive vision system much quicker than had it taken over 200 milliseconds to receive the audio feedback.

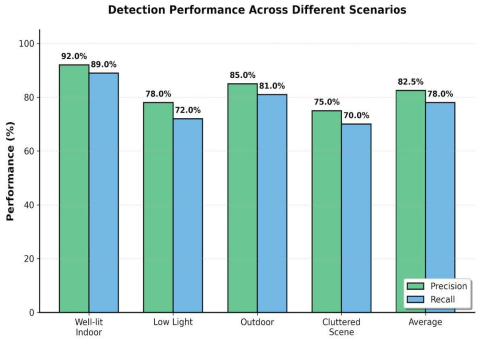

Fig 3. Performance Evaluation

Finally, the device used only 3.2W (watts) of power from the maximum available power for the device is 5W. This device is an excellent example of energy efficient, as the assistive vision system will be used by visually impaired users every day and use very little amount of energy as it will provide daily assistance for visually impaired people and will not require large amounts of energy to function.

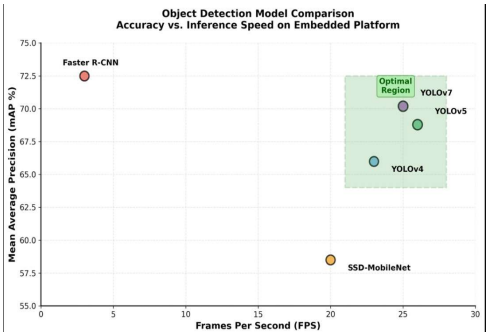

Fig 4. Accuracy Vs Inference Speed

In testing under varied real-world conditions and environments, there is confidence in the continued performance of the system as it is tested. The highest accuracy and recall results were achieved indoors with good lighting conditions, at 92% and 89% respectively, confirming that the established accuracy levels prior to testing continued to be valid. However, when testing in difficult environmental conditions, the system produced outputs, but the precision and recall accuracy for those outputs were lower than expected. In low-light environments, for example, the lowest precision score achieved was 78%, while precision remained relatively good (85%) when tested outside. However, when both of these sites had high degrees of clutter, or where multiple overlapping items were present (e.g., fallen tree branches), there will be a corresponding degree of drop-off in precision (i.e., approximately 75% precision). The average score for both tests was 82.5% (precision and recall); therefore, the solution appears to be viable for assisting visually impaired persons to safely navigate during their daily navigation tasks.

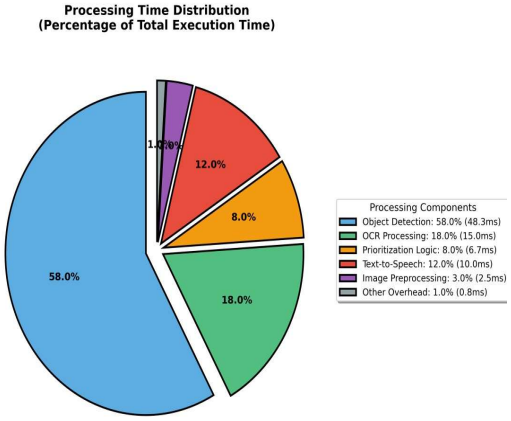

Fig 5. Processing Time Distribution

In our evaluation of the embedded system's performance for YOLOv5 as part of our comparative benchmark analysis, YOLOv5 is at the “optimal performance” level of performance for this application. Even though Faster R-CNN achieves the highest overall accuracy (72.5%) mAP (mean Average Precision), its 3 FPS frame rate is not usable for the user in real-time assistance applications. On the other hand, SSD-MobileNet can achieve a frame rate of 20 FPS; however, its mean Average Precision (mAP) of 58.5% is of little benefit to a user. Therefore, YOLOv5 has achieved an average mAP of 68.8% and operates at a frame rate of 26 FPS, providing the best performance of any algorithm tested, while also just slightly exceeding YOLOv4 (23 FPS) and being close to but slightly below YOLOv7 alternative for achieving mAP accuracy. YOLOv5 therefore provides us with a position in the identified “optimal” area, reinforcing the architectural decision of using YOLOv5 as the solution to providing real-time feedback to a person who is blind and/or visually impaired to locate objects in their physical space.

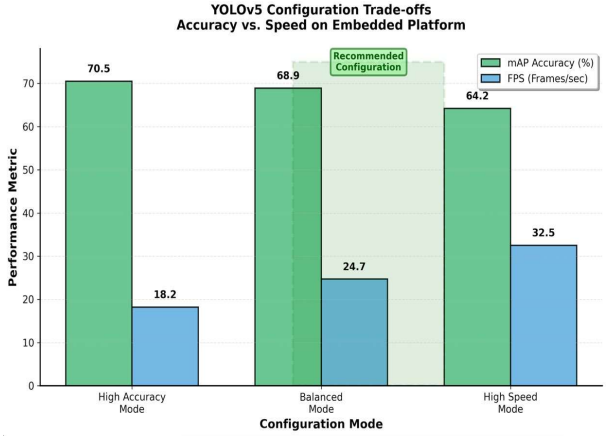

Fig 6. Configuration Trade-offs with various modes

Most of the processing time for all pipeline tasks was spent on object detection, which accounted for 58% (48.3 ms) of the processing time. The next most considerable computational workload was OCR, which used 18% (15 ms) of total processing time. Text to speech used the least amount (12%) at 10 ms. The prioritization logic consumed 8% of (6.7 ms) processing time in filtering and ranking detected objects based on their relevance to the user. In comparison, preprocessing images and system overhead consumed minimal processing time. Preprocessing images used 3% of total processing time, while system overhead was only 1%. This distribution supports a rationale for distributing resources effectively; 58% of all resources were used to complete the first task, which provides timely environmental awareness for the visually impaired.

V. Conclusion

Text-based object detection using YOLO v5 was evaluated for visually impaired individuals by recording results from a variety of configuration settings to determine performance metrics (accuracy). Recording made will facilitate future improvements based on the accuracy and deficiencies discovered through inference observations recorded during testing performed by using YOLO v5 compared to traditional methods. Additional research will occur to further enhance the system's real-time ability by incorporating an appropriate hardware platform.

References

[1] World Health Organization, World Report on Vision. Geneva, Switzerland: WHO Press, 2019.

[2] G. I. Okolo, T. Althobaiti, and N. Ramzan, "Assistive navigation systems for visually impaired persons using deep learning: A review," Sensors, vol. 24, no. 11, Art. no. 3572, 2024.

[3] M. Joshna, P. Ramesh, and K. S. Rao, "Object detection and audio feedback system for visually impaired persons," in Proc. Int. Conf. Recent Developments in Intelligent Computing and Communication Technologies (ICRDICCT), 2023, pp. 45–49.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[5] G. Jocher, A. Chaurasia, and J. Qiu, "YOLOv5," Ultralytics, 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[6] H. Luo, J. Wei, Y. Wang, J. Chen, and W. Li, "An improved lightweight object detection algorithm based on YOLOv5," PeerJ Computer Science, vol. 10, 2024.

[7] T. Mahendrakar, "Performance study of YOLOv5 and Faster R-CNN for autonomous navigation around non-cooperative targets," arXiv preprint arXiv:2301.09056, 2023.

[8] S. Tang, S. Zhang, and Y. Fang, "HIC-YOLOv5: Improved YOLOv5 for small object detection," IEEE Access, vol. 11, pp. 112340–112352, 2023.

[9] M. Wang, Z. Chen, and Y. Li, "FE-YOLOv5: Feature enhancement network based on YOLOv5 for small object detection," J. Visual Communication and Image Representation, vol. 90, 2023.

[10] S. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[11] W. Liu, D. Anguelov, D. Erhan, et al., "SSD: Single shot multibox detector," in Proc. European Conf. Computer Vision (ECCV), 2016, pp. 21–37.

[12] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[13] S. Mittal, "A survey of techniques for optimizing deep learning on edge devices," ACM Computing Surveys, vol. 54, no. 4, 2022.

[14] R. Smith, "An overview of the Tesseract OCR engine," in Proc. Int. Conf. Document Analysis and Recognition (ICDAR), 2007, pp. 629–633.

[15] P. Kumar and R. Singh, "Real-time text-to-speech system using OCR for visually impaired persons," Int. J. Advanced Computer Science and Applications, vol. 12, no. 8, 2021.