Designing an Assistive Tool for Visually Impaired People Based on Object Detection Technique

[1] Ghazwan Jabbar Ahmed *, [2] Farah Hatem Khorsheed *, [3] Fadhil Kadhim Zaidan

[1,2,3]University of Diyala, Diyala, Iraq

*Coresponding author.
E-mail address: farah_hatam@uodiyala.edu.iq

Abstract. *Visually impaired individuals often face significant challenges in navigating their environments due to limited access to visual information. To address this issue, we propose an assistive tool designed to operate on a PC. The focus of this research is on developing an efficient, lightweight object detection system to ensure real-time performance while maintaining compatibility with low-resource setups. The proposed system enhances the autonomy and accessibility of visually impaired individuals by providing audio descriptions of their surroundings through the processing of live-streaming video. The core of the system is an object detection module based on the state-of-the-art YOLO7 model, designed to identify multiple objects in real-time within the user's environment. The system processes video frames captured by a camera, identifies objects, and delivers the results as audio descriptions using the pyttsx3 text-to-speech library, ensuring offline functionality and robust performance. The system demonstrates satisfactory results, achieving inference speeds ranging from 0.12 to 1.14 seconds for object detection, as evaluated through quantitative metrics and subjective assessments. In conclusion, the proposed tool effectively aids visually impaired individuals by providing accurate and timely audio descriptions, thereby promoting greater independence and accessibility.*

Keywords: *Deep learning, YOLO, Object detection, Visually Impaired, Text-to-speech.*

INTRODUCTION

Visual data has now become very important in our daily lives. This importance comes from the impact of visual information on human thinking and its role in decision-making. The degree of visual impairment varies between individuals, with some people losing their ability to see completely [1], others only being able to distinguish light, shapes, or having no visual perception at all. The challenges faced by people with visual impairment vary depending on the degree to which their visual abilities are affected. According to the World Health Organization's 2017 estimates, there were approximately 253 million people with visual impairment, of whom 36 million were completely blind [2]. We can define visual impairment as a decrease in the ability to see, which in turn leads to vision problems that can never be solved with traditional solutions such as glasses and others. Therefore, it has become necessary to improve the quality of life for people with visual impairment.

In the past, people with visual impairment were helped with different techniques, as white canes help their users feel the way and enable them to reach some places through trained dogs. Despite the benefit and magnificence of these inventions for their users, they are classified as limited, as in order to benefit from the canes, they must be physically in contact with things all the time. The same is true for dogs. In order to benefit from them, they must be cared for and trained intensively. It must be noted that technological developments such as the global positioning system (GPS) and three-dimensional sound systems have played a major and influential role in improving the daily lives of visually impaired people over the world. However, the above-mentioned technologies have limited functions and focus on basic things such as knowing the measurement of angles [3,4] highlighting the urgent need for comprehensive assistive technology that can help users reasonably and understand the environment more deeply.

With the development of deep learning algorithms, technologies such as object detection [5], video translation [6], and image translation [7] have become essential tools to enhance accessibility for people with visual impairments after translating visual information into speech. Object detection helps people with visual impairments navigate their surroundings with confidence and safety by identifying real-time objects and potential hazards and by describing objects in daily life such as food, tables, chairs, etc. Object detection greatly helps people with visual impairments move independently in their environment. The most common algorithms for object detection are R-CNN, region-based convolutional neural network [8], SSD (single-shot multi-detector) [9], and also YOLO You Only Look Once [10]and its extensions are a two-stage detector It starts with creating a region proposal and then follows it by classifying these regions to identify the objects on them. R-CNN improved the process by using a region proposal network to create proposals with less computational time. This leads to an increase in the speed and accuracy of object detection. While SSD and YOLO can detect in one step, they are combined into one process, suggesting the region and classifying the object. This feature of integration gives them high detection speeds with less computational cost. This approach allows YOLO to achieve real-time object detection speed while maintaining reasonable accuracy. SSD is known for its balance between accuracy and speed.

However, YOLO's real-time performance may encourage researchers to release several versions of the model. For example, YOLO5 is known for its speed and accuracy. It has been used and achieved success in computer vision applications such as autonomous vehicles [11], video surveillance [12], and drone navigation [13]. Based on this, YOLO7 [14] was released, which provides significant improvements in speed and accuracy. It also features the detection of multiple objects in a single video or image frame, which could make it an important option for complex object detection [15,16]. It has been shown that YOLO7 [17] is more accurate than Faster-R-CNN and also has better real-time performance. This paper presents an assistive tool for visually impaired individuals, utilizing YOLO7 for object detection, the system delivers audio descriptions of objects for the user to facilitate a richer understanding of the environment using a camera and headphones. The key contributions of this work include:

1- Designing and Implementing an assistive system for the visually impaired utilizing the YOLO7 model.
2- Performing comprehensive experiments and comparisons across various versions of YOLO7 to assess speed, accuracy, and computational efficiency.
3- Designing a lightweight and cost-effective framework that integrates YOLO7 with real-time audio feedback, ensuring seamless operation on

embedded platforms for enhanced portability and usability.

A. Related work

The recent literature related to addressing problems related to computer vision was emphasized in Section 1, and its aim is to present appropriate methods to help people with visual impairment. V. Kumar et al [1] presented a method through which the blind can be helped by relying on image translation technology. The encryption can be decoded and a coding framework can be created by using ResNet50-LSTM networks. The study in [3] used a different technique to describe visual content in images based on VGG16-LSTM. With all of the above, both methods need an attention layer in their deep structure, which is very important for processing sequential data such as video and texts. The solution to this problem was found by M. Sarkar et al. [18], where the image captioning method was adopted, which is based on the concept of deep learning, which includes an attention mechanism. This deep model relies on the Inception-ResNet network, which was previously trained to extract features, followed by the gated (GRU) network to generate explanatory comments. In the study [19] the researchers described a single video frame through image captioning technology, which can help the blind, as this model can help one frame for every 50 frames in order to reduce complexity. In addition, the authors Included a method that can measure the distance of objects detected by the camera using Yolo5 and using the similarity of triangles approach. The captioning was based on a traditional structure using VGG16 [20] to extract features and using LSTM [21] to generate words. However, by integrating the attention mechanism into the deep neural structure, this model can be upgraded to perform better.

The study in [22] developed a system that can describe video to help visually impaired and blind people. This system integrates multiple pre-trained models as in Yolo3 to detect objects and a pre-trained image translation model. However, the system was primarily designed to process video frames from a pre-recorded media file instead of processing them directly from the camera, which limits its real-time applications. In the study [23], a deep learning model was proposed that can help blind people recognize information in the environment by using video translation techniques. A system was built by using an encoding and decoding framework using a VGG16 network and a group to decode captioning. using RNN-LSTM. Then, the MSVD dataset [24] was allocated into five categories to train the model. The proposal, however, lacks a mechanism for interest in the system, in addition to the use of traditional models.

B. Proposed Approach

Figure 1 illustrates the system architecture and the interaction between its hardware components. The hardware components include a PC (Personal Computer), a USB camera, and headphones. The camera captures video frames, which are processed by the object detection model. The results are then converted from text to audio format using the pyttsx3 Python library and delivered to the user via the headphones. Unlike online text-to-speech tools, pyttsx3 provides offline functionality, ensuring reliable performance without needing an internet connection.

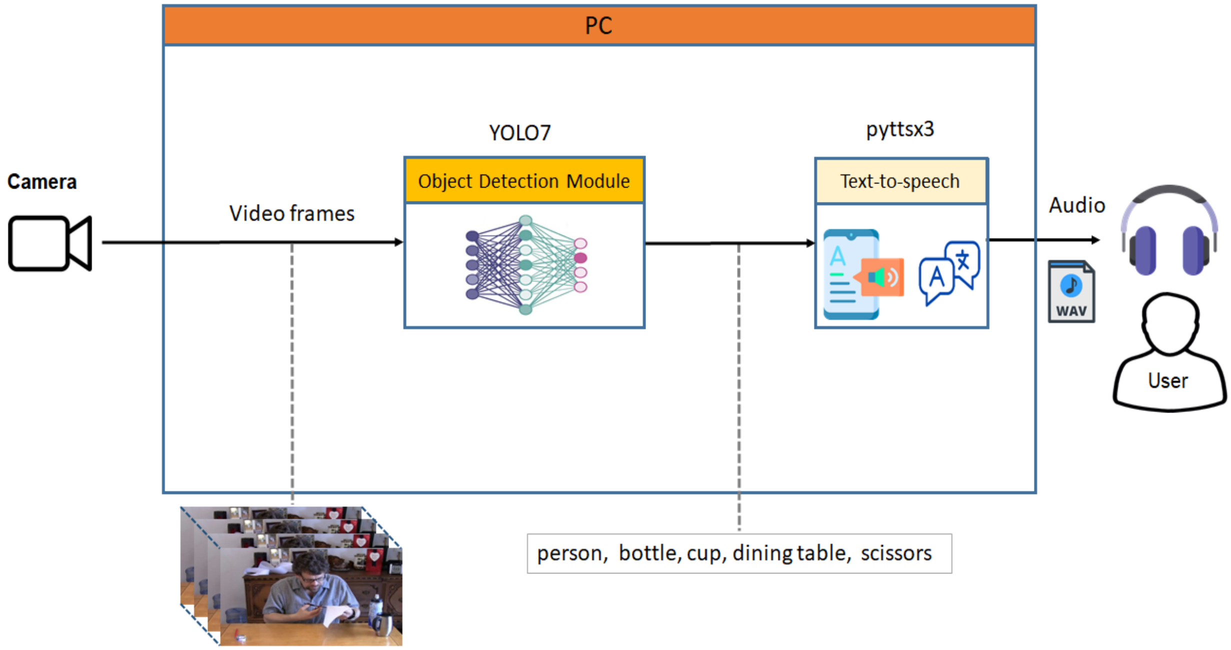

Figure 1. Architecture of the proposed system

In the proposed system, objects are detected using pre-trained Yolo7 as a basic framework based on transfer learning. Yolo7 is known for its high accuracy and efficiency in real-time object detection tasks. Here, we used Yolo7 as it is suitable for real-time applications on embedded systems. This model identifies objects within the frame and classifies them with special tags such as a car, a table, etc. After that, it ensures that the tags are suitable for speech synthesis by processing them. Finally, the processed tags are translated into audio descriptions using the text-to-speech Python library.

C. Experimental Results and Discussion

The object detection model is analyzed and evaluated in this section. The pre-trained YOLO7 is used, which has been trained on the well-known MS COCO dataset. In order to ensure efficiency, this system processes video frames at a rate of one frame per second. The frames that have been taken from the YOLO7 model are applied to detect objects on these frames. There is more than one version within the YOLO7 model, between YOLO7-tiny and the largest YOLO7-E6E. See Table 1, taking into account the comparison between speed and accuracy. Additionally, Figure 2 illustrates the qualitative results from real-world scenarios using various versions of YOLO7. This Figure highlights the ability of these models to detect multiple objects in real-time with high accuracy. However, as demonstrated in Figure 2, YOLO7-E6E achieves higher accuracy than others. The experiments are executed on a computer with a Core i7 processor and 8 GB of RAM using the Python programming language utilizing the Pytorch framework.

Table 1: Comparison of YOLO7 variants

Version	Number of parameters	Size	AP (evaluated on MS COCO)	Inference speed
YOLO-tiny	6.2M	12.3 MB	38.7%	0.12 s
YOLOv7	36.9M	73.8 MB	51.4%	0.42 s
YOLO-X	71.3M	139.7 MB	53.1%	0.84 s
YOLO-W6	70.4M	137.9 MB	54.9%	0.55 s
YOLO-E6	97.2M	190.4 MB	56.0%	0.70 s
YOLO-D6	133.7M	261.9 MB	56.6%	0.98 s
YOLO-E6E	151.7M	297.2 MB	56.8%	1.14 s

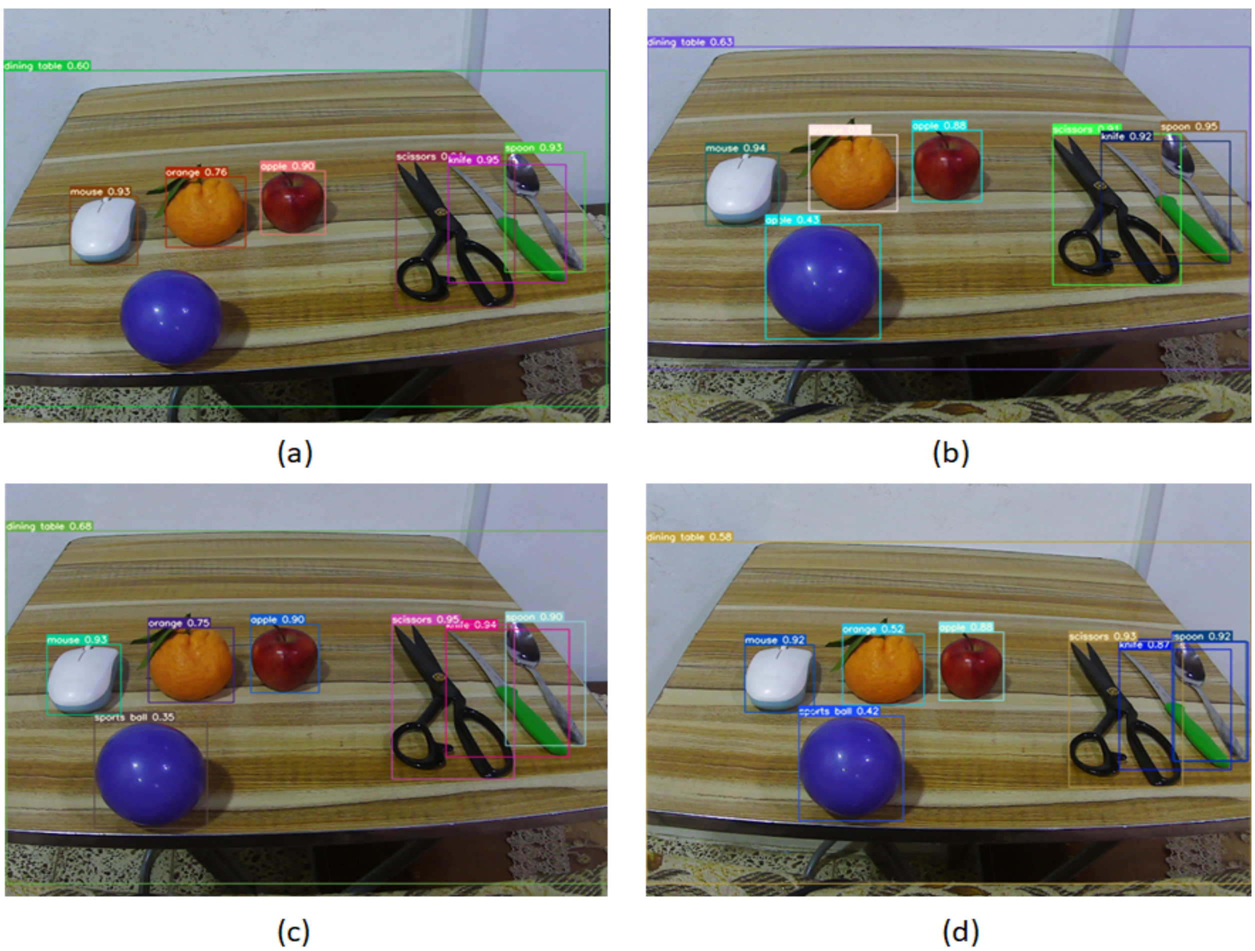

Figure 2. Result of Object detection using: (a) YOLO7-tiny, (b) YOLO7, (c) YOLO7-D6, (d) YOLO-E6E.

CONCLUSION

Using the best techniques available in deep learning, visually impaired individuals were helped through the hardware implementation system, which is what the study presented in this research. We also focus on the literature and analysis of related studies, highlighting their limitations and preparing for progress in this field. Our approach is mainly based on providing useful audio descriptions of the environment through video translation and object detection, to help and improve users' ability to move independently. Here we show that our system shows a significant development in accuracy and efficiency. Additionally, the proposed system does not require special skills to operate and is easy to set up, in addition to being cost-effective without needing for internet connection. Our experiments have proven the effectiveness of the system across different scenarios and its ability to adapt.

As for future work, we plan to deploy the proposed system on portable device like the Jetson Nano. Furthermore, the proposed system can be enhanced by adding other assistive technologies to it, such as text recognition, image captioning, and face recognition.

REFERENCES

[1] V. Kumar, V. Teja, A. Kumar, V. Harshavardhan and U. Sahith, "Image Summarizer for the Visually Impaired Using Deep Learning," 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), Puducherry, India, 2021, pp. 1-4, doi: 10.1109/ICSCAN53069.2021.9526465.

[2] B. Arystanbekov, A. Kuzdeuov, S. Nurgaliyev and H. A. Varol, "Image Captioning for the Visually Impaired and Blind: A Recipe for Low-Resource Languages," 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, 2023, pp. 1-4, doi: 10.1109/EMBC40787.2023.10340575.

[3] A. Chharia and R. Upadhyay, "Deep Recurrent Architecture based Scene Description Generator for Visually Impaired," 2020 12th International Congress on Ultra-Modern Telecommunications and Control Systems and Workshops (ICUMT), Brno, Czech Republic, 2020, pp. 136-141, doi: 10.1109/ICUMT51630.2020.9222441.

[4] C. C et al., "Image/Video Summarization in Text/Speech for Visually Impaired People," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972653.

[5] T. Mohandoss, J. Rangaraj, Multi-Object Detection using Enhanced YOLOv2 and LuNet Algorithms in Surveillance Videos, e-Prime - Advances in Electrical Engineering, Electronics and Energy, Volume 8, 2024, 100535, ISSN 2772-6711, https://doi.org/10.1016/j.prime.2024.100535.

[6] M. Sarkar, S. Biswas and B. Ganguly, "A Hybrid Transfer Learning Architecture Based Image Captioning Model for Assisting Visually Impaired," 2023 IEEE 3rd Applied Signal Processing Conference (ASPCON), India, 2023, pp. 211-215, doi: 10.1109/ASPCON59071.2023.10396262.

[7] A. Yousif and M. Al-Jammas, “Exploring deep learning approaches for video captioning: A comprehensive review,” e-Prime - Adv. Electr. Eng. Electron. Energy, vol. 6, no. October, p.

100372, 2023, doi: 10.1016/j.prime.2023.100372.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 580-587, doi: 10.1109/CVPR.2014.81.

[9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single Shot MultiBox Detector. In European conference on computer vision, pages 21–37. Springer, 2016.

[10] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[11] B. Xiao, J. Guo, and Z. He, "Real-Time Object Detection Algorithm of Autonomous Vehicles Based on Improved YOLOv5s," 2021 5th CAA International Conference on Vehicular Control and Intelligence (CVCI), Tianjin, China, 2021, pp. 1-6, doi: 10.1109/CVCI54083.2021.9661149.

[12] P. Zhang, W. Hou, D. Wu, B. Ge, L. Zhang, and H. Li, "Real-Time Detection of Small Targets for Video Surveillance Based on MS-YOLOv5," 2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 2023, pp. 690-694, doi: 10.1109/ICAIBD57115.2023.10206275.

[13] Y. Yang, "Drone-View Object Detection Based on the Improved YOLOv5," 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 2022, pp. 612-617, doi: 10.1109/EEBDA53927.2022.9744741.

[14] Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7464-7475. 2023.

[15] S. Chourasia, R. Bhojane and L. Heda, "Safety Helmet Detection: A Comparative Analysis Using YOLOv4, YOLOv5, and YOLOv7," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-8, doi: 10.1109/ICONAT57137.2023.10080723.

[16] T. Reddy Konala, A. Nammi and D. Sree Tella, "Analysis of Live Video Object Detection using YOLOv5 and YOLOv7," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-6, doi: 10.1109/INCET57972.2023.10169926.

[17] I. Hilali, A. Alfazi, N. Arfaoui and R. Ejbali, "Tourist Mobility Patterns: Faster R-CNN Versus YOLOv7 for Places of Interest Detection," in IEEE Access, vol. 11, pp. 130144-130154, 2023, doi: 10.1109/ACCESS.2023.3334633.

[18] M. Sarkar, S. Biswas and B. Ganguly, "A Hybrid Transfer Learning Architecture Based Image Captioning Model for Assisting Visually Impaired," 2023 IEEE 3rd Applied Signal Processing Conference (ASPCON), India, 2023, pp. 211-215, doi: 10.1109/ASPCON59071.2023.10396262.

[19] A. S. Alva, R. Nayana, N. Raza, G. S. Sampatrao and K. B. S. Reddy, "Object Detection and Video Analyser for the Visually Impaired," 2023 Third International Conference on Artificial

Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 1405-1412, doi: 10.1109/ICAIS56108.2023.10073662.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[21] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Compute. 9 (1997) 1735–1780.

[22] A. Bodi, P. Fazli, S. Ihorn, Y. Siu, A. Scott, L. Narins, Y. Kant, A. Das, and I. Yoon. 2021. Automated Video Description for Blind and Low Vision Users. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 1–7. https://doi.org/10.1145/3411763.3451810.

[23] Y. -H. Huang and Y. -Z. Hsieh, "The Assisted Environment Information for Blind based on Video Captioning Method," 2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), Taoyuan, Taiwan, 2020, pp. 1-2, doi: 10.1109/ICCE-Taiwan49838.2020.9258088.

[24] D. Chen and W. Dolan, "Collecting highly parallel data for paraphrase evaluation". In ACL: Human Language Technologies- Volume 1. ACL, 190-200, 2011.

Conflict of Interest Statement:

The author declares that the research was conducted in the absence of any commercial or financial relation- ships that could be construed as a potential conflict of interest.

Article History:

Received: 7 August 2025 | Accepted: 9 August 2025| Published: 30 November 2025

.

Copyright © 2025 Ghazwan Jabbar Ahmed , Farah Hatem Khorsheed , Fadhil Kadhim Zaidan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms