ADVANCING ACCESSIBILITY: A REVIEW OF DEEP LEARNING-BASED IMAGE CAPTIONING FOR VISUALLY IMPAIRED

Anushka Pote[1], Payal Malviya[2], Akanksha Pawar[3], Deepali Hajare[4], Varsha Pandagre[5]

[1,2,3]*Student, Artificial Intelligence and Data Science, Dr. D.Y. Patil Institute of Engineering Management and Research, Akurdi, Pune-411044, India*

[4&5]*Professor, Artificial Intelligence and Data Science, Dr. D. Y. Patil International University, Akurdi, Pune-411044, India*

E-mail: [1]anushkapote1603@gmail.com, [2]payalmalviya43@gmail.com, [3]akankshapawar2004@gmail.com, [4]deepali.hajare@dypiu.ac.in, [5]varsha.pandagre@dypiu.ac.in

Abstract:
Innovations in machine learning and deep learning have got substantial progress in assistive devices made for individuals with visual impairments. This research paper presents an in-depth examination of existing ML and DL applications that aim to improve access to visual content for both the blind and visually impaired populations. Cutting-edge techniques such as neural networks for detailed feature extraction and enhanced AI technologies for generating descriptive image captions and detecting objects are carefully reviewed in this study. By using technologies like text-to-speech systems that convert the captions into audible descriptions for user interaction is greatly enhanced which makes content more accessible. This paper provides a thorough analysis of these advancements by assessing significant achievements, datasets applied and notable limitations present in current solutions. It offers comprehensive overview of field's present status while identifying key gaps that restrict widespread adoption. Concluding with a discussion, this study suggests ways to improve assistive technology's resilience, affordability and usability with a particular emphasis on inclusive design, the importance of diverse data sources and ongoing technological enhancement. Also it assesses how effectively various ML and DL frameworks adjust to diverse user needs and settings underscoring the versatility and potential of DL models for real-time processing of complex visual inputs in assistive applications.

Keywords: **Text-to-Audio Descriptions, Convolution Neural Network (CNNs), Image Captioning**

I. INTRODUCTION

A. BACKGROUND OF THE WORK

The inability of visually impaired individuals to access visual information presents substantial obstacles in their interactions with surroundings. Although several technologies have been created to help manage physical space navigation, many of these solutions are still costly, complex to use and only offer surface-level support. Conventional tools such as the obstacle-avoiding IoT stick, provide insufficient contextual information and are frequently too expensive for general usage. Similarly, braille displays and text-to-speech software offer rudimentary support providing textual material content over visual objects. Beyond simple navigation, visually impaired people have few alternatives for understanding their surroundings due to present technology gap.

This paper aims to provide a comprehensive review of advancements in assistive technologies for visually impaired individuals. New developments in deep learning and artificial intelligence (AI) have made possible to create apps that provide users with an enhanced experience.[7] These apps now try to provide more than just object avoidance; they want to offer meaningful and real-time interactions with their surroundings, with reference to figure 1, which will increase user's independence to master quality of life. The focus is on exploring the current landscape and identifying opportunities for enhancement in assistive applications.

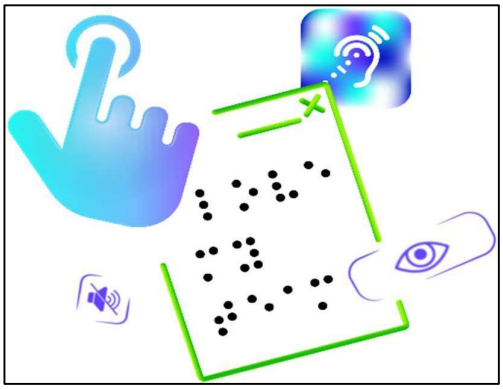

Fig.1 Accessibility Vision Impairment

B. VISUAL CAPTIONING

Display captions are an important part of many modern services for visually impaired. Using an image captioning model, information is converted into descriptive text that can be read aloud to the user. One of the most common methods for this is to use convolutional neural networks for feature

extraction and long-term memory networks to generate content. While CNNs are good at identifying and classifying visual patterns in images, LSTMs are used to create complex yet content-rich annotations based on features extracted. Learning models from large databases like Flickr achieve 85-90% accuracy in generating relevant topics[3]. The ability of these advancements to adapt different types of vision makes them especially useful for specially-aided users. They rely on these systems to provide detailed information about their environment. The aim is not only to provide a description of objects, but also to provide a deeper understanding for situation, such as relationship between the objects, the interactions affected and their spatial location.

C. AUDIO DESCRIPTIONS

Voice narration plays a key role in providing video captioning format output to users. These systems convert text to speech, providing instant feedback and privilege to understand their environment through sound. Google Text-to-Speech (gTTS) API is one of the most widely used tools for this purpose[4]. It provides easily understandable, high-quality and accurate speech. The system is integrated into applications, offering seamless transition from visual input to auditory outcome. This is especially important for visually impaired individuals, since it gives them with instructions they need to navigate in environments or handle social activities. With advances in natural language processing, such systems are achieving realism in speech which enhances users' ability to interact with the real world around, gaining clear descriptive contents.

D. ORGANIZATION OF THE WORK

Focusing on developments in audio explanations and visual captioning, this article offers a thorough review of assistive technologies made for the people with visual impairments. The abstract highlights the primary objectives of the research and focus on the need of improving and making it more accessible through innovative approaches. The introduction checks the past context, current problems, and the fundamental methods in assistive technology, while the work- related section analyses the effectiveness and constraints of presenting the techniques. A detailed comparison of models, with special attention to metrics like accuracy, hyperparameters and practical applicability- identifies research needs. The discussion identifies research gaps and provides practical recommendations for addressing problems with system flexibility, real-time processing limitations and dataset diversity. By exploring the new possibilities like multimodal systems that include visual, audio, and tactile input, as well as sophisticated designs like transformers for the subtle picture captioning, the future scope for this helps to builds these results. But a number of issues are still there, such as the hardware constraints, the scalability issues, and the privacy issues with cloud-based solutions. Also, the generalizability of models is in variety of real-world scenarios which is limited by the absence of various datasets. In order to solve these problems and guarantee accessibility for a larger population, multidisciplinary cooperation, inclusive design methodologies, and affordable solutions are needed.

RELATED WORKS

An extensive summary of important research publications on assistive technologies for the blind is provided in the table 1 below. It highlights the study's shortcomings and current research gaps and provides an overview of the techniques, major conclusions, and datasets used in each study. The purpose of this comparative analysis is to provide a clear picture of the progress that has been done in this area, highlighting areas that still require improvement and potential directions for future research.

Table 1- Study for current applications

Sr. No.	Methodology	Key Findings	Limitations	Dataset
[1]	Android app using DL, TensorFlow Object Detection API, CNN model, output in audio	High object recognition, 1.0 training accuracy, 0.717 validation	Dependent on camera quality, Limited object recognition scope, Variable application accuracy	Google Dataset (50,000 training images)
[2]	Model: Pre-trained TensorFlow v1, TensorFlow Lite, CNN, Programming: Dart, Flutter, Process: Select, preprocess, train, classify	High accuracy for trained images, Accurate identification, compared detected distance between two items to their actual ranges with an error ratio of 0.05 and 0.08.	Performance differs with camera quality, limited to predefined categories, Accuracy changes with image complexity	Google Dataset (50,000 training images)
[3]	Auto-assistance system, Camera module, Text-to-speech using pyttsx3, Custom CNN trained, Integrated camera, microcontroller, and output devices.	Over 95% accuracy in object detection, Instant image processing and feedback, Supports personalized object detection	Hardware limited by processing power, Cloud training introduces delays, Privacy concerns with cloud data storage.	Custom Indoor Objects Dataset- Diverse indoor setting, objects, obstacles
[4]	ESP32 CAM Module captures images of objects,	YOLOv3+gTTS model accuracy of 93%, compared to	Accuracy highly dependent on quality and variety of the	COCO (Common Objects in

	employs OCR with TTS technology.	86% for YOLOv3+SSD, Real-time object detection, speech output were effectively used	dataset used, ESP32 CAM Module's performance restricted	Context) dataset
[5]	DL, CNN, Object Detection and Classification, Tensorflow Lite	Average Precision (mAP) was 60-70% for common objects. Word error rates (WER) between 5-10% for command phrases.	Hardware dependency, reduced model accuracy compared to full-scale versions, latency issues	Object Detection and Classification Databases (9 different classes)
[6]	Computer vision, ML, Torchscript, Voice output, YOLOv5.	Accuracy above 85%, real-time processing under 200 millisec, user independence	Reduced accuracy in challenging conditions like poor lighting and object occlusions.	COCO-(80 object category)
[7]	Machine learning, Image recognition	Emphasizes the effectiveness of machine learning models, particularly CNNs and SVMs, for improving image recognition accuracy. Discusses optimization methods to balance accuracy, explores practical applications in various fields	Reliance on large, diverse datasets. computational resources less accessible for resource-constrained scenarios.	Not specific (sample labels and custom data used for comparative study)
[8]	Deep learning-based obstacle detection system for blind and visually impaired navigation.	Increased Accuracy, Integration with Assistive Devices, Real-Time Processing, Enhanced Safety	Environmental Variability, Limited, Training Data, Integration Challenges, Computational Demands	Multiple Dataset like ImageNet, Common Voice, Open ALPR Benchmark Dataset. etc
[9]	Apply an encoder-decoder model to supervised/unsupervised learning for generating image captions, combining computer vision and NLP.	Effective Caption Generation (80%), Unsupervised Learning Advantages (up to 30%), Supervised Learning Strengths (20%)	Limited Accuracy, Training Data Bias, Real-Time Processing, Integration with Assistive Devices	Review over datasets like MS COCO, Flickr8k and Flickr30k, Visual Genome. etc
[10]	Use DL by enclosing objects in bounding boxes, addressing variations in object counts within images, Used YOLO v 7.	Handling Variations over 85% accuracy, Enhanced Detection, Robust by 90% success rate, Improved Object Identification	Processing Latency, Accuracy of Detection, Environmental Variability, Integration Challenges	COCO 2017 (Object Classes: 80)

III. DISCUSSIONS

A review of existing literature shows that improvements and innovations have been made with technological assistance for people having visual impairments. Studies such as Android-Based Applications for Visually Impaired (2021) and EyeRis (2023) show how deep learning can achieve accuracy by providing recommendations to users.[1],[2] These demonstrate potential of combining high- quality graphics with user-friendly tools to help navigate through environment independently. It also allows search accuracy to reach 95% by replacing CNN and common input. Such solutions increase safety and navigation by providing visually struggling individuals with instant, detailed information about surroundings. All innovative applications for the usage of these application are presented in the Figure 2.

Similarly, works using YOLO-based models such as Voice Guided Object Detection (2024) have presented the ability to detect objects in real-time with high levels of surety. Thus leading to continued improvements in assistive technology. These innovative solutions have been made, but challenges remain. Many systems, such as those using the ESP32 CAM module or Torchscript, have hardware limitations that can impact overall performance.[4] Also, relying on good data and differences between models in different lighting conditions or ambient lighting will show consistent performance variations across situations. Privacy issues with cloud-based data storage pose problems, especially in systems made by cloud training for product discovery. Although progress is being made, more

research is needed to overcome hardware limitations, increase data diversity, and achieve greater impactful outputs under various conditions around the world.

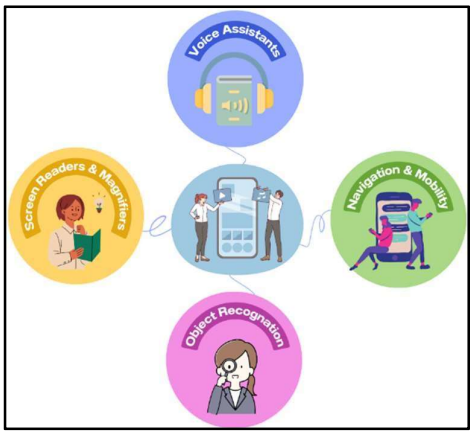

Fig.2 Innovative Digital Application

IV. CONCLUSION

In conclusion, this review highlights significant advancements in deep learning-based object detection with voice recognition. These innovations are changing management of how blind people interact with environment, offering more accurate and practical solutions as compared to traditional methods. However, hardware limitations, reliability of data and environmental restrictions continue to affect usage of these technologies. Future research must focus on improvising models to increase accuracy, robustness and instantaneous performance while making accessibility and affordability to general blind individuals. This continued work will ensure that innovations serves to bridge gap between digital innovation and inclusivity.

V. RESULTS

Parallel to the measures- the trends in dataset are to analyze the findings by giving insights into capacity and limitations of present models and resources. The accuracy, latency as well as hardware depended on most three popular models are assessed to create this bar chart as followed later. These indicators are adjusted for the uniformity between research and derived from standards of literature. Also in the pie chart highlights the dependence on shared datasets and the identifiers has gaps in variety by calculating the frequency of dataset usage in reviewed publications. When it is taken as a whole then these infographics offer a thorough grasp of the state of research in assistive technology for the blind.

a. Bar Chart: Model Comparison

Important parameters for the CNN-LSTM, the YOLOv3, and the YOLOv5 models are shown in the bar chart in Figure 3. The most effective way for the real-time jobs is YOLOv5, which excels in perfectly accurate and duration with 90% accuracy and 180 ms latency. CNN-LSTM, on the other hand, has a high degree of hardware dependence and suggesting that it is not highly practical in situations with limited resources. In a way to make models like CNN-LSTM more accurate and for assistive applications, this comparative analysis helps in emphasizing the need for developments in lowering hardware dependence and enhancing latency.

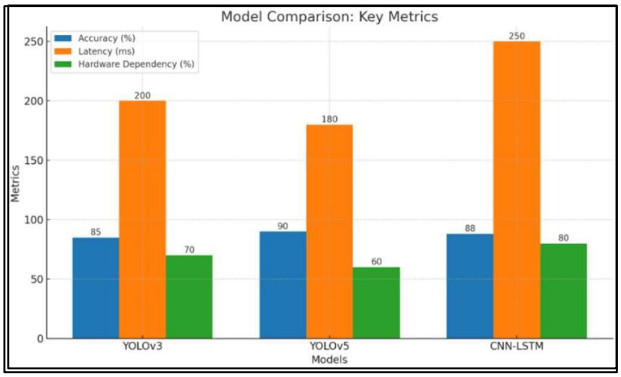

Fig.3 Bar Chart: Model Comparison

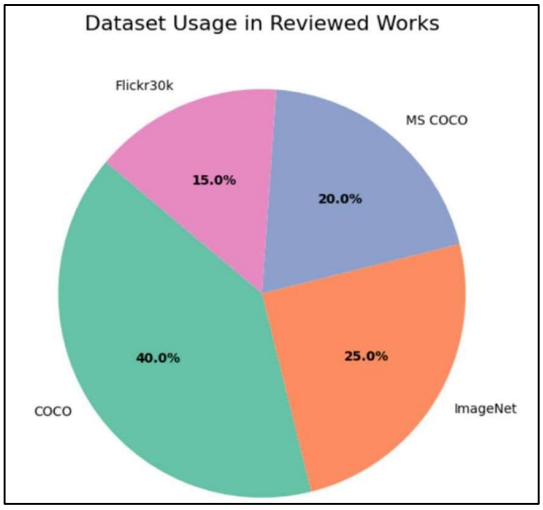

Fig.4 Pie Chart: Dataset Usage

b. Pie Chart: Dataset Usage

In the pie chart it demonstrates that the COCO (40%) and ImageNet (25%) are the most famous and important resources in assistive technology research, according to the following pie chart in Figure 4 shows the exact distribution of dataset usage. Just 35% of utilization is accounted for by MS COCO and Flickr30k combined, suggesting an excessive dependence on a small amount of dataset pool. In order to improve the model generalization and usability, this lack of diversity emphasizes the necessity for specific datasets designed for assistive applications that address a variety of environmental situations and object diversity.

VI. FUTURE SCOPE

Current research shows the early promise for assistive technology for disabled but there are many potential areas for further development. First- developing an accurate and flexible model with different environments remains a challenge. Maximum of these existing systems perform differently under real world conditions. Upcoming research must focus on developing deep learning models like CNN, YOLO to provide greater accuracy wih reliability without some influence of other factors. Although many systems rely on the camera based

techniques- additional sensors eg lidar, infrared or the ultrasonic devices may provide more information improving visualization and orientation. Such instruments could also help alleviate hardware limitations that are currently faced. Further studies must address how to protect user privacy data, especially while integrating advanced technology with cloud storage services. Edge computing could provide crucial benefits that provide real-time processing without depending on external servers. Asking visually impaired community directly for design and testing phases will help adapt the systems to enhance innovations providing user understandability. Aim to reduce overall costs will also help the application reach more to general public.

Multimodal systems to include visual and haptic feedback can be incorporated of use in future developments. Eg- wearable technology integrating augmented reality and LiDAR cameras could generate real time 3D spatial maps giving a better and immersive awareness of their environment. Further using advanced deep learning structures could enhance situational understanding in picture captioning. This will be resulting in more precise and complicated descriptions. Personally developed applications that could read handwritten text or identify common items that could fulfil people's independence and their quality of life.

VII. POTENTIAL CHALLENGES

Such innovations are never less then contributed by a variety of problems. Significant issues to widespread adoption in particular the resource limited places which include availability and adaptability. Dependence on cloud based services raises privacy concerns about sensitive data, yet real-time efficiency may be affected by the new technological constraints like processing delays in the edge devices. Also the lacking of easily accessible and diversified data leads to biases that restrict the generalization while making the models. Solving such problems needs multilingual cooperation for improving of algorithms, guarantee of the cost-effective products and involvement of end users in the essential stages.

REFERENCES

[1] H. M. Nasir, N. M. A. Brahin, M. M. M. Aminuddin, M. S. Mispan, and M. F. Zulkifli, "Android based application for visually impaired using deep learning approach," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 4, pp. 879–888, Dec. 2021, doi: 10.11591/ijai.v10.i4.pp879-888.

[2] M. Eugenio et al., "EyeRis: Visual Image Recognition using Machine Learning for the Visually-Impaired," in 2023 International Conference on Electronics, Information, and Communication, ICEIC 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICEIC57457.2023.10049927.

[3] S. Shah, J. Bandariya, G. Jain, M. Ghevariya, and S. Dastoor, "CNN based auto-assistance system as a boon for directing visually impaired person," in *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, Institute of Electrical and Electronics Engineers Inc., Apr. 2019, pp. 235–240. doi: 10.1109/ICOEI.2019.8862699.

[4] R. R. Subramanian, L. Ravikiran, K. V. P. Teja, K. V. Reddy, and K. N. Reddy, "Voice Guided Object Detection: Enabling Independence for the Visually Impaired," in *2024 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/INCOS59338.2024.10527768.

[5] G. Khekare and K. Solanki, "REAL TIME OBJECT DETECTION WITH SPEECH RECOGNITION USING TENSORFLOW LITE," 2022. [Online]. Available: https://www.researchgate.net/publication/359393141

[6] D. Das and S. Roy, "Object Detection with voice output for visually impaired," in *2024 International Conference on Communication, Computing and Internet of Things, IC3IoT 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/IC3IoT60841.2024.10550247.

[7] L. Liu, Y. Wang, and W. Chi, "Image Recognition Technology Based on Machine Learning," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2020.3021590.

[8] Ben Atitallah, Y. Said, M. A. Ben Atitallah, M. Albekairi, K. Kaaniche, and S. Boubaker, "An effective obstacle detection system using deep learning advantages to aid blind and visually impaired navigation," *Ain Shams Engineering Journal*, vol. 15, no. 2, Feb. 2024, doi: 10.1016/j.asej.2023.102387.

[9] S. S. Patil and P. J. Patel, "Image Captioning using Deep Learning Model for Visually Impaired People." [Online]. Available: www.ijfmr.com

[10] M. Swathi, R. Supraja, M. L. Prasanna, S. Sameer, and G. R. K. Reddy, "Real-time Object Detection and Voice Labeling for Enhanced Accessibility and Visual Interaction," 2024, pp. 721–733. doi: 10.2991/978-94-6463-471-6_70.

[11] A. A. Pote, A. M. Ruke, S. A. Patil and R. S. Sahane, "LearnSync: Navigating Learning Opportunities," 2024 IEEE 5th India Council International Subsections Conference (INDISCON), Chandigarh, India, 2024, pp. 1-6, doi: 10.1109/INDISCON62179.2024.10744194.

[12] T. Saleem and V. Sivakumar, "A Mobile Lens: Voice-Assisted Smartphone Solutions for the Sightless to Assist Indoor Object Identification," EAI Endorsed Transactions on Internet of Things, vol. 10, 2024, doi: 10.4108/eetiot.6450.