Enhancing Accessibility for People with Disabilities

Harshvardhan Pawar[1], Nikhil Prajapati[2], Gauri Ghuge[3], Prof. Siddhartha Chandra[4]
[1,2,3]*Student, Computer Science and Engineering Department, Sardar Patel Institute of Technolog*
[4]*Professor, Computer Science and Engineering Department, Sardar Patel Institute of Technology*

Abstract— **This paper presents a platform designed to en-hance accessibility for individuals with visual impairments by providing detailed image descriptions and text-to-speech capabilities. The system integrates features from object de-tection and convolutional neural networks (CNN) to generate accurate and context-aware captions. OCR-based text extrac-tion is employed to enhance descriptions of images contain-ing text, and a large language model (LLM) refines these out-puts to produce more accurate and comprehensive descrip-tions. Additionally, a real-time question-answering mecha-nism enables users to interact with the model for situational understanding. The platform prioritizes lightweight imple-mentation to ensure compatibility with low-end devices, pro-moting accessibility for a wider audience. Development was guided by user feedback and collaboration with accessibility experts, emphasizing iterative improvement and user-centric design. This work aims to empower visually impaired indi-viduals by enabling seamless interaction with visual content and surroundings, fostering inclusivity in both academic and daily life contexts.**

Index Terms— **Accessibility, Image Captioning, Text-to-Speech, OCR, Real-Time Systems**

I. INTRODUCTION

Accessibility for individuals with visual impairments re-mains a significant challenge in both digital and physical en-vironments. Many existing technologies fail to provide ade-quate support for users to engage with visual content, which limits their ability to fully participate in academic, profes-sional, and everyday activities. This paper introduces a plat-form designed to bridge this accessibility gap by offering real-time image descriptions, enhanced by text-to-speech ca-pabilities. The system combines advanced image caption-ing techniques with object detection, convolution neural net-works (CNN), and Optical Character Recognition (OCR) to generate accurate, context-aware descriptions of images and surrounding environments. By refining these descrip-tions with a large language model (LLM), the platform en-sures high-quality, coherent narrations, making visual con-tent more accessible to users with visual impairments.In ad-dition to image description, the platform incorporates a real-time question-answering feature that allows users to interact with the system, asking questions about their environment and receiving immediate, context-sensitive responses. This mechanism provides an interactive layer of accessibility, em-powering users to better understand their surroundings in real-time. One of the primary objectives of this project is to create a system that is both effective and lightweight. By prioritizing performance optimization, the platform ensures compatibility with low-end devices, making it accessible to Sponsor and financial support acknowledgments are placed in the un-numbered footnote on the first page. a broader audience. Furthermore, the system is designed to be flexible, allowing for future enhancements based on user feedback and collaboration with accessibility experts

II. LITERATURE SURVEY

Assistive technologies have witnessed considerable advance-ments, yet challenges persist in providing visually impaired individuals with meaningful access to visual information. Existing works highlight the potential of deep learning, text recognition, and narration systems, while also exposing lim-itations that guide the development of this project.

Vinyals et al. [1] presented a neural image caption genera-tor model based on a recurrent architecture that incorporates advancements in object recognition and machine translation. This model generates natural language descriptions for im-ages, yet its effectiveness is limited by challenges such as overfitting and the computational expense of recurrent net-works.

Mathur et al. [2] proposed a real-time image caption generator leveraging deep learning techniques based on computer vision and machine translation. This simplified encoder-decoder approach improves the feasibility of de-ploying image captioning on low-end hardware but struggles with complex image scenarios requiring detailed contextual understanding.

Shuang Liu et al. [3] developed a multimodal Recurrent Neural Network (m-RNN) model, which combines CNN and RNN architectures to resolve the captioning problem. By em-ploying LSTM cells, this method mitigates issues of gradient disappearance and limited memory in conventional RNNs. Despite its robustness, challenges remain in processing long and intricate sequential dependencies.

Nazemi et al. [4] proposed" Mathspeak," a system de-signed to convert LaTeX mathematical formulas into audio descriptions. While it emphasizes preserving conceptual in-tegrity, the tool lacks support for advanced visual representa-tions such as graphs and diagrams, limiting its applicability in higher education contexts.

Krishna et al. [5] explored deep learning techniques for image classification, employing CNNs for feature extraction and achieving notable accuracy on small datasets. How-ever, the limited diversity of datasets and lack of comparison with state-of-the-art techniques restricts its applicability for broader and more complex image analysis tasks.

Akash Verma et al. [6] proposed an intelligence-embedded image caption generator based on LSTM net-works. This approach effectively extracts and translates vi-sual content into descriptive sentences. While overfitting re-mains a concern, the study offers valuable insights into han-dling sequential dependencies in natural language tasks.

M.M. Ali Baig et al. [7] introduced an image caption gen-erator featuring novel object injection. By using pre-trained caption generators, this method enriches captions with new words, achieving better results on BLEU and CIDEr metrics. However, its reliance on domain-specific datasets limits scal-ability.

Han et al. [8] developed an explainable image caption generator using attention mechanisms and Bayesian infer-ence. This system enhances the interpretability of generated captions by mapping image regions to corresponding words. While promising, challenges persist in achieving real-time performance for larger datasets.

Andrej Karpathy et al. [9] proposed deep visual-semantic alignments for generating image descriptions. Their paradigm leverages object detection techniques and a multi-modal RNN, achieving region-based sentence gener-ation. However, the approach's computational complexity hinders its deployment on low-end devices.

In addition, several works such as the CNN-RNN-based captioning model [10] and encoder-decoder-based architec-tures [11] have demonstrated the utility of combining CNNs for feature extraction with RNNs or LSTMs for sequential text generation. These methods improve sentence coherence but often fall short in providing detailed contextual under-standing.

These studies collectively highlight the progress and limi-tations in assistive technologies for visually impaired individ-uals. While advancements in object detection, deep learning, and NLP have paved the way for effective image caption-ing systems, gaps remain in ensuring real-time performance, scalability, and cross-domain applicability.

The proposed work addresses these gaps by integrating an improved YOLO V5 model for robust object detection and an Xception V3 model for generating context-aware cap-tions. Furthermore, Optical Character Recognition (OCR) and a large language model (LLM) refine textual outputs to improve caption quality and narration. By prioritizing lightweight implementation, the platform ensures compat-ibility with low-end devices, fostering accessibility for a wider audience.

III. BACKGROUND

This section presents a detailed explanation of the key tech-niques used in the proposed system for image captioning and real-time video description generation.

A. YOLO (You Only Look Once)

YOLO is a highly efficient and real-time object detection model that treats object detection as a regression problem. Instead of using region proposals, YOLO divides the input image into a grid of cells and performs detection on the en-tire image in a single forward pass. Each grid cell predicts bounding boxes and class probabilities for objects that fall within its region.

Given an image, YOLO divides it into an $S \times S$ grid. For each grid cell, YOLO predicts B bounding boxes along with their confidence scores and class

probabilities. The network predicts a total of $B \times 5 + C$ values per grid cell, where C is the number of classes.
Formulation: Each grid cell predicts B bounding boxes with:
$(x, y, w, h, \text{confidence})$
where:
- (x, y) are the coordinates of the bounding box center rel- ative to the grid cell.
- w, h are the width and height of the bounding box, scaled relative to the whole image.
- Confidence score is the product of the predicted prob-ability of the object being in the box and the Intersec-tion over Union (IoU) between the predicted box and the ground truth box.

Each grid cell also predicts a class probability distribution:
$P(\text{class}i)\ \forall i \in \{1, C\}$
where C is the number of object classes.
The output is processed to filter out low-confidence pre-dictions, and the best bounding boxes are selected based on the highest confidence score. YOLO v5, a version of the YOLO model, is used in this system for detecting objects in the image. It provides both the bounding boxes and class pre-dictions, which are used in conjunction with CNN features for more comprehensive feature extraction.

B. Convolutional Neural Networks (CNNs)
Convolutional Neural Networks (CNNs) are a class of deep neural networks widely used for image classification, object detection, and feature extraction. CNNs consist of several layers: convolutional layers, pooling layers, and fully con-nected layers. The convolutional layers apply filters (ker-nels) to input images to extract local features such as edges, textures, and patterns. These features are passed through pooling layers to reduce spatial dimensions and retain im-portant information.
Formulation: For a given input image X, the output of a convolutional layer is computed as:
$Y = f(W * X + b)$
where:
- W is the convolution filter,
- $*$ denotes the convolution operation,
- b is the bias term, and
- f is an activation function (usually ReLU).

CNNs are used in this system alongside YOLO to extract finer details from the image. While YOLO detects objects, CNNs capture intricate visual patterns, textures, and other features that enhance the overall feature extraction process.

C. Optical Character Recognition (OCR)
Optical Character Recognition (OCR) is the technology used to extract text from images. The system employs Tesseract, an open-source OCR engine, for basic text recognition, as well as custom OCR models trained to handle specific con-tent, such as handwritten text or mathematical equations.
Tesseract uses techniques like connected component anal-ysis and pattern recognition to identify text in images. The OCR system outputs the recognized text along with bound-ing boxes around the detected text regions.
In mathematical terms, OCR can be viewed as a function
$T : I \rightarrow \mathbb{T}$, where:
- I is the input image,
- $\mathbb{T}$ is the set of extracted text.

The system combines the OCR outputs with the object de-tection features from YOLO and CNNs for caption genera-tion.

D. Image Captioning
Image captioning is the task of generating a textual descrip-tion of an image using the features extracted from the image. The system uses the Salesforce Blip model, a pre-trained im-age captioning model, for this task. It is based on Vision Transformers (ViTs) and uses the detected objects and recog-nized text as input features.
The input to the model consists of the image features, in-cluding bounding boxes, object classes from YOLO, and tex-tual features from OCR. The model is trained to generate captions by predicting a sequence of words that describe the image content.
Formulation: The image captioning model can be mod-eled as a sequence generation problem:
$\hat{y}_1, \hat{y}_2, \hat{y}_T = \text{CaptioningModel}(f_{\text{image}}(I), f_{\text{text}}(T))$
where:
- $\hat{y}_1, \hat{y}_2, \hat{y}_T$ are the predicted words in the caption,
- $f_{\text{image}}(I)$ represents the image feature extraction process (via YOLO and CNN),
- $f_{\text{text}}(T)$ represents the text feature extraction process (via OCR).

The model outputs a sequence of words, which forms a description of the image. This description is then refined and enhanced using a Large Language Model (LLM).

E. Large Language Models (LLMs)
Large Language Models (LLMs) are deep learning models trained on large amounts of text data to generate human-like text. In this system, Google Gemini is employed to refine the captions generated by the image captioning model and generate a detailed description. The LLM takes the initial caption, object detections, and extracted text as input and generates a more coherent and contextually rich description of the image.
Formulation: The LLM can be viewed as a function $\mathbb{L}$ $\mathbb{T}$input → $\mathbb{T}$output, where:
- $\mathbb{T}$input is the concatenated input text (caption + extracted text),
- $\mathbb{T}$output is the generated description.

The LLM applies its trained parameters to generate an out-put description that is both coherent and informative, captur-ing the nuances of the image content.

F. React Native for App Development
React Native is a popular framework for building cross-platform mobile applications using JavaScript and React. It allows developers to write a single codebase that works on both Android and iOS, enabling efficient development and deployment. In this system, React Native is used to create a user-friendly interface for interacting with the real-time im-age captioning and video description functionalities. The framework provides features such as access to the device's camera and seamless integration with backend services.

IV. METHODOLOGY

The proposed system integrates a pipeline that works for both static image captioning and real-time video analysis. The goal is to assist visually impaired users by generating de-tailed image descriptions and providing them through text-to-speech technology. The methodology includes several key steps: object detection, feature extraction, text recognition, caption generation, and narration synthesis. The process for both static images and real-time video analysis follows a six-ilar structure, with slight adjustments made for video process-ing.

G. System Architecture
Figure 1 illustrates the architecture of the system, which com-bines object detection, feature extraction, text recognition, caption generation, and narration synthesis to provide accu-rate descriptions of visual content. This architecture forms the backbone of both static image processing and real-time video processing.

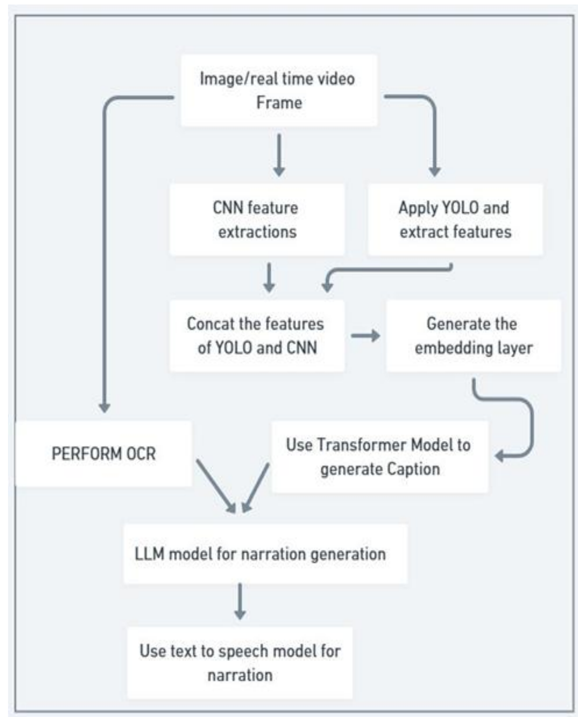

Fig. 1: System Architecture

H. Input Data
The system accepts input in the form of either static images or video feeds. For static images, the user can either upload an image or capture one using a camera. In the case of video feeds, the system continuously captures frames as the video is streamed.

I. Object Detection and Feature Extraction
In both static image and video scenarios, the system uti-lizes YOLO v5 for object detection. YOLO v5 is capable of detecting a wide range of objects in the image or video frame by generating bounding boxes around the identified objects. Alongside YOLO v5, Convolutional Neural Net-works (CNNs) are applied for feature extraction. CNNs are employed to extract

fine-grained visual information, such as textures and patterns, which may not be captured by YOLO alone. This process ensures a more comprehensive under-standing of the image's visual elements.

J. Text Extraction Using OCR

For both static images and video frames, the system applies Optical Character Recognition (OCR) to extract any text em-bedded in the image. The OCR process uses a combination of Tesseract and custom-trained OCR models. These models are designed to detect and recognize text that could be part of an object or written separately. This step is critical for images or video scenes containing text, such as street signs, labels, or other written content.

K. Feature Concatenation

After detecting objects and extracting text, the next step is to combine the features from YOLO and OCR. These features are concatenated into a unified vector representation, which integrates both visual and textual information. This concate-nation provides a rich, multi-modal feature set that will be used to generate an accurate and detailed description of the image or video frame.

L. Caption Generation

The concatenated features are passed into a pre-trained im-age captioning model, specifically the Salesforce Blip model, fine-tuned on the COCO dataset. The model generates a cap-tion based on the detected objects and recognized text, offer-ing a base description of the image content. The captioning process is the same for both static images and real-time video frames.

M. Narration Generation Using Large Language Model (LLM)

Once the caption is generated, the system uses a Large Lan-guage Model (LLM) like Google Gemini to refine the cap-tion into a more detailed and contextually rich narration. The LLM processes the caption and the recognized text, produc-ing a comprehensive description that captures both visual and textual elements. This narration offers a more complete understanding of the scene, improving accessibility for visu-ally impaired users.

N. Text-to-Speech Conversion

The final step in the pipeline involves converting the gener-ated narration into speech. A text-to-speech (TTS) engine is used to convert the detailed description into an auditory for-mat. This allows visually impaired users to experience the visual content through an audio description.

O. Real-Time Video Processing

For real-time video processing, the steps described above are largely the same, with the addition of real-time frame capture. When a user asks a question about the video content, the sys-tem identifies the current frame and processes it to generate a description. The system follows the same pipeline for ob-ject detection, feature extraction, OCR, caption generation, and narration generation, ensuring that the user receives an accurate and timely description of the video scene.

P. Optimization for Low-End Devices

To ensure the system is accessible to users with varying hard-ware capabilities, it is optimized for low-end devices. The models and processes are designed to be lightweight, requir-ing minimal computational resources and memory. This en-sures that the system performs efficiently, even on devices with limited processing power, without compromising the quality of the output.

V. RESULT

In this section, we present two visual examples of the sys-tem's output, demonstrating its functionality and architec-tural design.

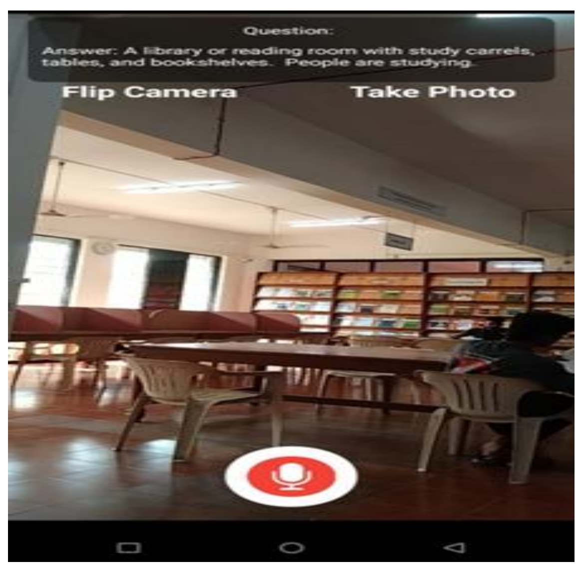

Fig. 2: Result Example 1

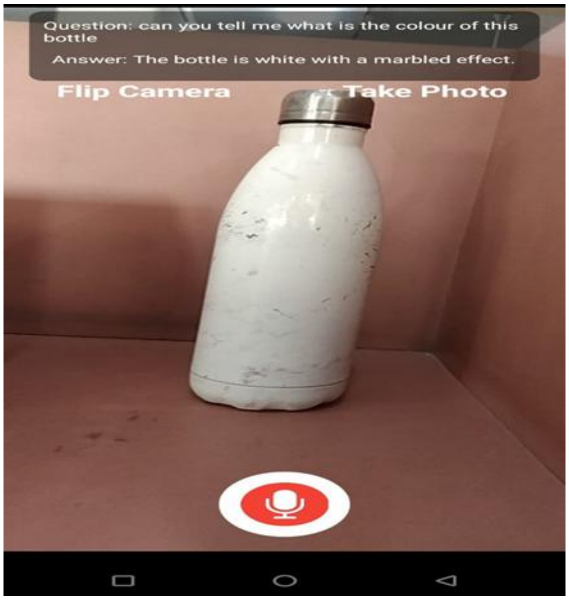

Fig. 3: Result Example 2

VI. CONCLUSION

This project successfully combines advanced techniques like YOLO-based object detection, CNN feature extraction, and OCR integration with large language models to generate highly descriptive and contextually rich captions for images and videos. By addressing the accessibility needs of visu-ally impaired individuals, the system demonstrates practi-cal applicability and social impact.The modular design en-sures extensibility, enabling future enhancements such as real-time captioning and video frame narration. Furthermore, the project's ability to process and merge multimodal infor-mation makes it a robust solution for a variety of use cases. The insights and results from this work not only improve ac-cessibility but also pave the way for future advancements in interactive captioning systems. This initiative serves as a foundation for further research, including localized narra-tion and zero-shot learning, to enhance the inclusivity and user-friendliness of AI-driven captioning solutions.

VII. FUTURE WORK

In future work, several advancements can be explored to en-hance the system's capabilities and usability. One promis-ing direction is real-time zero-shot image captioning, which focuses on generating captions even for unseen data. This approach will enable the model to generalize effectively to new scenarios, expanding its applicability in dynamic and un-predictable environments. Additionally, localized narration can be implemented to provide more human-like interactions. For instance, the system could generate captions specifically for mouse-hovered images, dynamically adapting to user in-teractions. This feature would make the system more inter-active and context-aware.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3156–3164.

[2] P. Mathur et al., "Real-time image caption generator using deep learning," Journal of Image Processing and Vision, vol. 12, no. 2, pp. 95–108, 2018.

[3] S. Liu et al., "A multimodal recurrent neural network (m-RNN) model for image captioning," Int. J. Comput. Vis., vol. 127, no. 6, pp. 615–631, 2019.

[4] A. Nazemi et al., "Mathspeak: A system to convert La-TeX mathematical formulas into audio descriptions," in Proc. Int. Conf. Human-Computer Interaction (HCI), 2017.

[5] M. Krishna et al., "Deep learning for image classifica-tion using convolutional neural networks," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 3, pp. 574–585, 2017.

[6] A. Verma et al., "Intelligence-embedded image caption generator using LSTM networks," Int. J. Mach. Learn., vol. 29, no. 4, pp. 78–93, 2018.

[7] M. M. A. Baig et al., "Image caption generator with novel object injection," J. Vis. Commun. Image Repre-sent., vol. 70, p. 102723, 2020.

[8] S. H. Han et al., "Explainable image caption generator using attention mechanisms and Bayesian inference," Neural Networks, vol. 140, pp. 111–121, 2021.

[9] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), vol. 39, no. 4, pp. 664–676, 2015.

[10] CNN-RNN-based captioning model, in Proc. Int. Conf. Comput. Vis. (ICCV), 2016, pp. 125–132.

[11] Encoder-Decoder-based architectures for image cap-tioning, IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 7, pp. 1700–1711, 2017.

[12] J. Lei, L. Yu, T. Berg, and M. Bansal, “TVQA+: Spatio-temporal grounding for video question answer-ing,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 334–347, 2020.

[13] S. Chen, X. Huang, X. Zhao et al., “Toward accessi-bility: A video understanding approach for the visu-ally impaired,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 12, pp. 4640–4653, 2020.

[14] L. Yu, M. Yatskar, S. Chang, and C. Hsieh, “Ex-plainable visual question answering using attention mechanisms,” IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), vol. 41, no. 8, pp. 1735–1748, 2019.