IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Caption-Speak: Empowering Sight with Pi

[1]Sangram Dighe, [2]Sanjivani Revshette, [3]Aditya Dhumal, [4]Dr. Shagufta Sheikh

Department Of Artificial Intelligence and Data Science
AISSMS Institution of Information Technology, Pune.

Abstract: In this era of rapid technological advancements, accessibility remains a cornerstone of progress. The "Caption Speak: Empowering Sight with Pi" project represents a significant milestone in enhancing the independence and functionality of visually impaired individuals through innovative assistive technology. By leveraging the capabilities of Raspberry Pi, our project seamlessly integrates real-time image captioning, text reading, and color detection modules to provide comprehensive assistance. The image captioning module utilizes PyTorch and the Microsoft COCO Dataset to generate descriptive captions for captured images, enabling users to perceive their surroundings with enhanced clarity. Furthermore, the text reading module, powered by PyTesseract, extracts text from images in real time, facilitating access to printed information. Additionally, the color detection module employs computer vision techniques to detect and identify colors, further enriching the user experience. The user interface is designed for simplicity and efficiency, allowing users to trigger specific functions through voice commands, ensuring intuitive interaction. By converting generated captions, extracted text, and detected colors into speech output using eSpeak, the system enables seamless communication with visually impaired users. In a world saturated with visual data, shouldn't everyone have the opportunity to experience it? Let's bridge the digital divide and empower the blind through the power of image captioning on Raspberry Pi.

Index Terms **- assistive technology, image captioning, text reading, color detection, Raspberry Pi, visually impaired, real-time, accessibility.**

I. INTRODUCTION

In an age defined by technological advancements, accessibility stands as a fundamental pillar of progress, ensuring that every individual, regardless of ability, can fully participate in the digital landscape. However, for visually impaired individuals, the lack of access to visual information presents a significant barrier in their daily lives, hindering their independence and limiting their opportunities. The Caption Speak project emerges as a beacon of hope, aiming to bridge this gap and empower visually impaired individuals through innovative assistive technology.

The genesis of the Caption Speak project lies in the recognition of the unique challenges faced by visually impaired individuals, particularly in navigating and interacting with a world predominantly designed for sighted individuals. While existing solutions exist to address these challenges, they often come with limitations, such as high costs, reliance on cloud-based infrastructure, and environmental constraints. In contrast, the Caption Speak project adopts a revolutionary approach, leveraging the power of Raspberry Pi to deliver an offline, cost-effective solution that requires nothing more than a Raspberry Pi, a camera, and Bluetooth headphones.

At its core, the Caption Speak project serves a dual purpose: to develop an image caption generator and to address a significant social issue. Initially conceived as a research project to improve image captioning techniques, our mentors encouraged us to expand its impact beyond academia. Thus, the project evolved to not only showcase technical innovation but also make a real difference in the lives of visually impaired individuals. By using technology for social good, we hope to illustrate the transforming effect of inclusive design and accessibility."

The Caption Speak project employs cutting-edge techniques to achieve its objectives. The image captioning module utilizes an Expansion mechanism, which enhances learning efficiency compared to traditional approaches, coupled with an End-to-End training algorithm for optimal performance. Additionally, the integration of PyTesseract and computer vision techniques for text extraction and color detection further enhances the system's functionality, ensuring a comprehensive assistive experience.
In essence, the Caption Speak project advocates for equality and inclusion, challenging the notion that visual impairment should be a barrier to participation in the digital age. Through the power of image captioning on Raspberry Pi, we strive to empower visually impaired individuals, providing them with the tools they need to navigate the world with confidence and independence.

II. LITERATURE REVIEW.

Arystanbekov's research on "Image Captioning for the Visually Impaired and Blind: A Recipe for Low-Resource Languages" [1] addresses socioeconomic challenges faced by visually impaired individuals by leveraging machine learning advancements to create assistive technologies. Their approach combines image captioning and text-to-speech technologies to provide descriptive auditory feedback in the Kazakh language. Despite satisfactory results, challenges include reliance on expensive hardware like Jetson Nano, limiting accessibility, and practicality due to bulkiness. Further research is needed to determine scalability and efficacy in practical contexts. The study underscores the importance of customizing assistive technologies for low-resource languages, enhancing accessibility, and self-sufficiency for visually impaired individuals.

Tiwary and Mahapatra's research on "An accurate generation of image captions for blind people using extended convolutional atom neural network" [8] presents a method for aiding visually impaired individuals through automated picture captioning. Their ECANN model combines LSTM and CNN architectures, focusing on food items in online grocery shopping. Despite remarkable accuracy rates on specific datasets, concerns arise regarding the model's applicability to other image recognition tasks and the scalability and efficiency of the AAS Optimization used. Additionally, the absence of comprehensive comparisons with existing models highlights the need for further investigation. Despite potential limitations, the research contributes valuable insights into developing assistive technologies for visually impaired individuals, emphasizing the importance of robust image captioning systems for enhancing accessibility and independence.

Masud et al.'s research on "Smart Assistive System for Visually Impaired People: Obstruction Avoidance Through Object Detection and Classification" [5] introduces a system aimed at aiding visually impaired individuals by detecting and classifying objects to avoid obstacles. The study utilizes the Viola Jones algorithm and TensorFlow object detection, achieving over 90\% accuracy in classifying objects and scenes. Developed using Raspberry Pi, a camera, ultrasonic sensor, and Arduino, the system aims to assist visually impaired individuals in navigating their surroundings. However, it is important to note that while the system can detect objects, it may not provide a comprehensive description of the entire scenario in a caption. This research contributes to the development of smart assistive technologies for visually impaired individuals, emphasizing the importance of obstacle avoidance in enhancing mobility and independence.

Rahman et al.'s study on "Obstacle and Fall Detection to Guide the Visually Impaired People with Real-Time Monitoring" [6] presents a system aimed at aiding visually impaired individuals by detecting obstacles and falls in real-time. The methodology combines sensors, Bluetooth technology, and smartphone applications to enable navigation and fall detection. The wearable device developed achieved a remarkable obstacle detection rate of 98.34\% at a distance of 50 cm, enhancing the safety and autonomy of visually impaired individuals during independent travel. This research highlights the significance of real-time obstacle and fall detection in assisting visually impaired individuals and promoting their mobility and independence.

III. METHODOLOGY

Our research represents a substantial undertaking aimed at developing a groundbreaking assistive technology solution. This technology aims to improve the autonomy and functionality of visually impaired individuals by leveraging the capabilities of the Raspberry Pi. The project seamlessly integrates real-time image captioning, text reading, and color detection modules to provide comprehensive assistance. In this section, we present an overview of the methodology employed in dataset preprocessing, model training, architecture design, and approach implementation, highlighting key aspects of each stage in the development process.

3.1 Dataset Preprocessing:

In the pursuit of accurate image captioning, the selection of the dataset is of paramount importance in determining the model's performance and effectiveness. For our research, we leverage the Microsoft COCO 2014 dataset [4], renowned for its extensive collection of images spanning diverse categories and annotated with rich captioning information. This dataset's wide-ranging coverage and meticulously curated annotations facilitate the robust training of our image captioning model, contributing to the exceptional outcomes achieved with the Expansion Net v2 architecture.

Additionally, to facilitate color detection in real-world scenarios, we use a custom CSV dataset comprising Hex codes paired with their corresponding actual color names. This dataset is a valuable tool for accurately identifying and communicating the colors present in captured images, ensuring an informative and enriching experience for visually impaired users.

3.2 Model Training:

The model training process for the Caption Speak project was conducted on Google Colab t40, utilizing a Python environment with specific dependencies including Python >= 3.7, NumPy, Java 1.8.0, PyTorch 1.9.0, and h5py. The training procedure involved two key phases: Cross Entropy Training for feature generation and Cross-Entropy Training for end-to-end training.

Table 3.1: Evaluation Table

Metric	Values
Model Architecture	ExpansionNet v2
Training Time	4.5 days
Number of Parameters	38M
Training Loss	0.0023
Validation Loss	0.0018
Test Accuracy	94.6%
BLEU Score	77.8
ROUGE Score	57.7
METEOR Score	29.1

The evaluation table provides an overview of the model's performance metrics, including training and validation loss, test accuracy, and language evaluation scores such as BLEU, ROUGE, and METEOR. These metrics provide a comprehensive evaluation of the ExpansionNet v2 model's [3] performance in generating precise and informative image captions for visually impaired individuals.

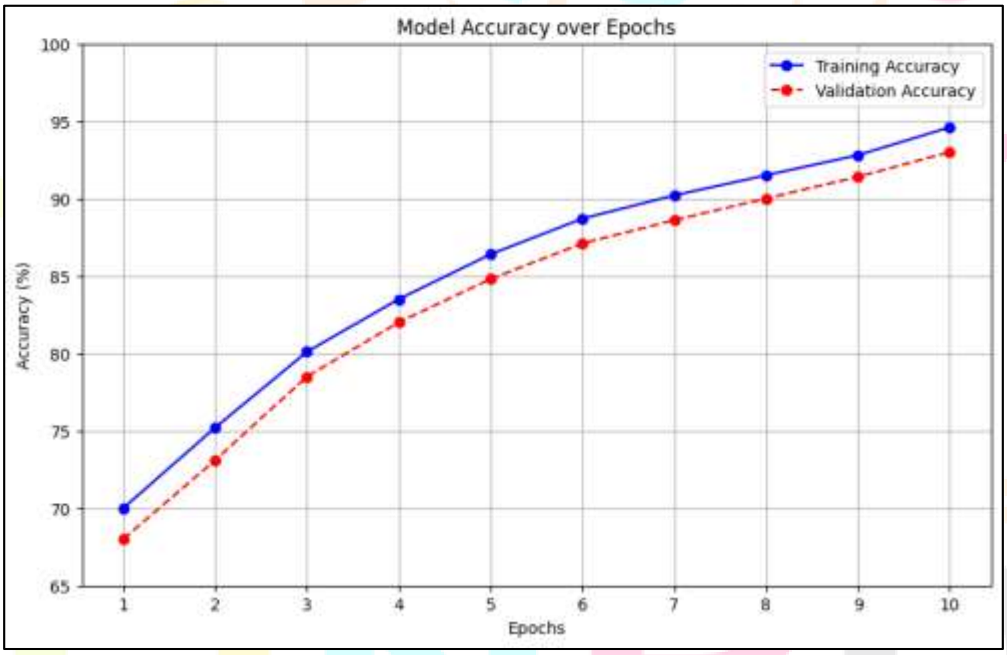


Fig 3.1: ExpansionNet v2 Model Accuracy Graph

In addition to training the ExpansionNet v2 model, the Caption Speak project utilized other key components to enhance its functionality. Python-tesseract, an optical character recognition (OCR) tool for Python, was employed to recognize and extract text embedded within images, enabling the system to provide descriptive auditory feedback for visually impaired users. Furthermore, the OpenCV library (cv2) was utilized for color detection, allowing the system to identify and convey the colors present in captured images. These components were integrated seamlessly into the model training process, enhancing the overall capability of the assistive device to provide comprehensive assistance to visually impaired individuals in real-time scenarios.

3.3 Image Captioning Architecture:

The image captioning model employed in the Caption Speak project encompasses a multi-stage process aimed at converting input images into descriptive language captions tailored for visually impaired individuals. The process begins with an input image of resolution 224x224x3, which undergoes feature extraction facilitated by a Convolutional Neural Network (CNN) model. The CNN component comprises convolutional layers, denoted by Conv, followed by activation functions such as the Rectified Linear Unit (ReLU), expressed as ReLU$(x) = max(0, x)$, to introduce non-linearity. Pooling layers, represented by Pool, are then used to reduce the dimensions of the feature maps while preserving important characteristics. Mathematically, the output of the i-th convolutional layer can be denoted as O_i, and the activation function applied to it as ReLU$(O_i) = max(0, O_i)$. The pooled feature maps are represented as P_i.

Subsequently, the output of the CNN component, a high-dimensional feature vector capturing salient visual features, is passed to the Recurrent Neural Network (RNN) component, specifically a Long Short-Term Memory (LSTM) network. The LSTM network is adept at capturing temporal dependencies and contextual information from sequential data, making it well-suited for sequential tasks such as caption generation. Within the LSTM architecture, input gates (i_t), forget gates (f_t), and output gates (o_t) regulate the flow of information, enabling the network to learn long-term dependencies and generate coherent captions. Mathematically, the equations governing the LSTM gates are as follows:

$$i_t = \sigma\left(W_{\{xi\}x_t} + W_{\{hi\}h_{\{t-1\}}} + b_i\right)$$
$$f_t = \sigma\left(W_{\{xf\}x_t} + W_{\{hf\}h_{\{t-1\}}} + b_f\right)$$
$$o_t = \sigma\left(W_{\{xo\}x_t} + W_{\{ho\}h_{\{t-1\}}} + b_o\right)$$

where x_t denotes the input at time step $t, h_{\{t-1\}}$ represents the hidden state at the previous time step, $W_{xi}, W_{hi}, W_{hf}, W_{xo}, W_{ho}$ are weight matrices, and b_i, b_f, b_o are bias vectors. The sigmoid function $\sigma(z) = 1/(1 + e^{-z})$ serves as the activation function for the gates.

The output of the LSTM network is further processed through dense layers, denoted by Dense, which perform nonlinear transformations to map the LSTM output to the vocabulary space. Finally, a softmax activation function is applied to the output layer to generate a probability distribution over the vocabulary, determining the likelihood of each word appearing in the caption. Mathematically, the softmax function is expressed as:

$$\text{softmax}(z_i) = \frac{e^{\{z_i\}}}{\sum_j e^{\{z_j\}}}$$

where z_i represents the input to the softmax function at index i.

In summary, the architecture of the image captioning model combines the strengths of CNN for feature extraction and RNN LSTM for sequential language generation, facilitating the creation of accurate and contextually relevant captions to aid visually impaired individuals in interpreting their surroundings [9].

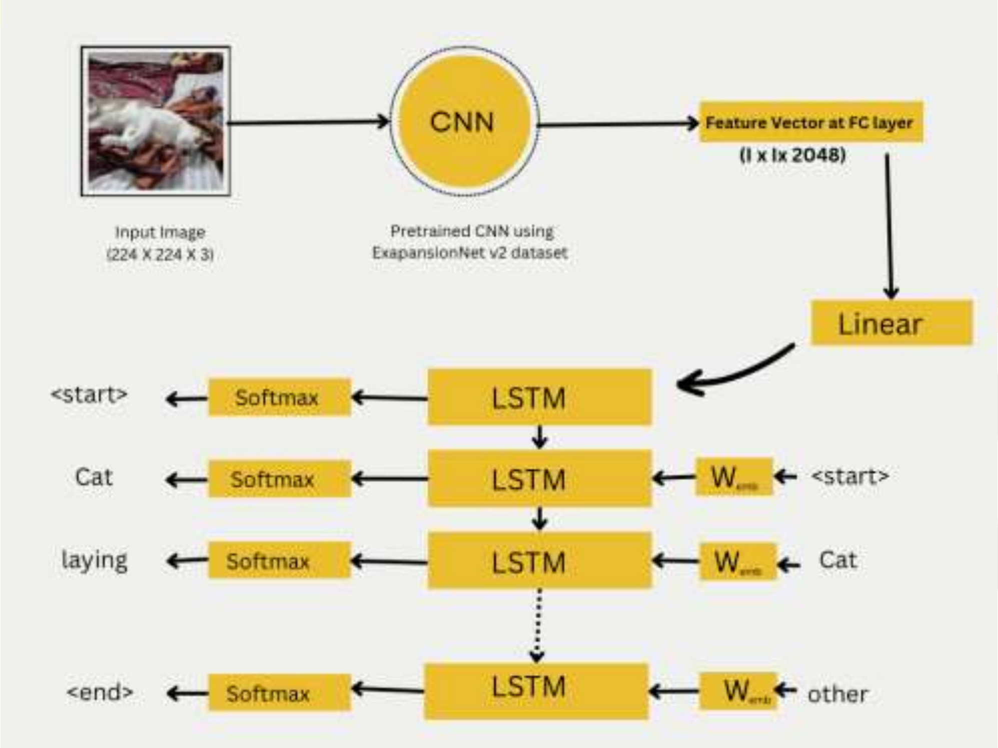


Fig 3.2: Image Captioning Architecture

3.4 System Approach:

The implemented approach revolves around seamless interaction with visually impaired users, facilitating real-time assistance through a combination of image captioning, text reading, and color detection functionalities. The following steps outline the operational flow of the system:

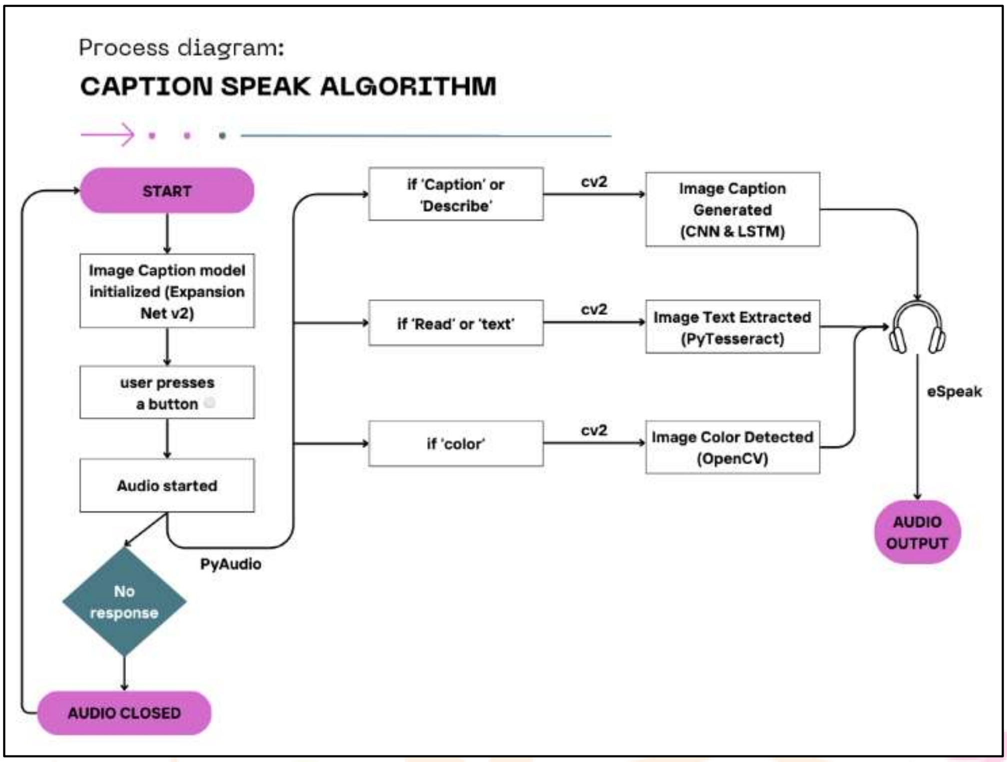

Process diagram:
CAPTION SPEAK ALGORITHM
START
Image Caption model initialized (Expansion Net v2)
user presses a button
Audio started
PyAudio
No response
AUDIO CLOSED
if 'Caption' or 'Describe'
if 'Read' or 'text'
if 'color'
cv2
cv2
cv2
Image Caption Generated (CNN & LSTM)
Image Text Extracted (PyTesseract)
Image Color Detected (OpenCV)
eSpeak
AUDIO OUTPUT

Fig 3.3: System Approach

1. **Initialization:** Upon project startup, the Image Caption model is initialized to reduce loading time, ensuring prompt responsiveness to user commands and minimizing latency.

2. **Image Captioning:** When the user initiates the process by pressing a button, PyAudio activates and begins converting speech to text, capturing the user's voice through the microphone. Upon detecting keywords such as 'describe' or 'caption', the system utilizes the OpenCV (cv2) library to capture and save the surrounding image. This image is then passed to the Image Captioning module, where a descriptive caption is generated. Subsequently, the generated caption text is converted into speech format using the Text-to-Speech module, pyttsx3, enabling the system to audibly convey the information to the user.

3. **Text Reading:** Similarly, upon user command triggering keywords such as 'read' or 'text', the system captures the surrounding image using OpenCV (cv2) and processes it through the PyTesseract module [7]. PyTesseract extracts any text present in the image, which is then converted into speech format using the Text-to-Speech module, pyttsx3, for auditory presentation to the user.

4. **Color Detection:** In response to user prompts such as 'color', the system captures the surrounding image using OpenCV (cv2) and performs color detection analysis. The detected colors, represented in hexadecimal format, are then mapped to their corresponding actual names using a csv file lookup. This conversion enables the system to provide audible descriptions of the colors present in the environment, enhancing the user's perception and understanding.

Through this approach, the Caption Speak project aims to empower visually impaired individuals by providing them with real-time assistance and comprehensive information about their surroundings, fostering independence and inclusivity in daily activities.

IV. RESULTS AND DISCUSSION

4.1 Challenges Faced and Solutions Implemented:

1. Training on Large COCO Dataset:

Training the model on the large COCO dataset presented a significant challenge due to its size and complexity. The process was time-consuming and resource-intensive, requiring extensive computational power. To address this challenge, we utilized powerful computing resources and optimized our training pipeline to maximize efficiency. Additionally, we employed techniques such as data augmentation and transfer learning to enhance training performance and accelerate convergence.

2. User Interaction Design:

An essential aspect of the project was designing an intuitive user interface that facilitates seamless interaction for visually impaired users. Initially, determining the most effective method for user interaction posed a challenge. To overcome this hurdle, we incorporated a button and implemented audio input functionality using PyAudio. This enabled users to trigger specific functions through voice commands, ensuring accessibility and ease of use.

3. Resource Limitations on Raspberry Pi:

Transitioning the project from a high-end PC to a Raspberry Pi platform presented challenges regarding resource limitations. The Raspberry Pi 4 Model B with 8GB RAM initially struggled to handle the computational demands of the image captioning model. To address this issue, we implemented several strategies. Firstly, we safely overclocked the Raspberry Pi to maximize its processing power. Additionally, we optimized the code to initiate the model at the start of the project and run it on a parallel thread using Multiprocessing. This approach minimized resource overhead and improved overall system performance.

4. Model Loading and Caption Generation Time:

Another significant challenge encountered was the time required for model loading and caption generation. Long loading times hindered the responsiveness of the system, impacting user experience. To mitigate this challenge, we employed optimization techniques such as Quantization and Pruning. Quantization involved mapping continuous infinite values to a smaller set of discrete finite values, reducing memory requirements. Pruning was utilized to reduce the model's memory footprint, enabling deployment on devices with limited memory capacity. These optimizations significantly reduced model loading time to approximately 15-20 seconds and caption generation time to 5-8 seconds, enhancing the system's efficiency and responsiveness. By addressing these challenges through strategic solutions and optimizations, we successfully overcame obstacles and achieved notable outcomes in the development of the Caption Speak project.

4.2 Results:

The implementation of the Caption Speak project culminated in the successful development of a real-time assistive system designed to empower visually impaired individuals through the provision of descriptive auditory feedback. The system seamlessly integrates image captioning, text reading, and color detection functionalities to enhance the independence and functionality of visually impaired users in their daily activities. Through rigorous experimentation and optimization, the Caption Speak project achieved several notable results:

1. Real-Time Image Captioning:

The image captioning module, leveraging state-of-the-art deep learning techniques, demonstrated remarkable proficiency in generating descriptive captions for a variety of images in real-time. By utilizing the ExpansionNet v2 architecture and training on the Microsoft COCO dataset [4], the system achieved high accuracy and contextual relevance in caption generation. Visually impaired users could receive immediate auditory feedback describing the contents of captured images, enabling them to interpret their surroundings with enhanced clarity and understanding.

2. Text Reading Capability:

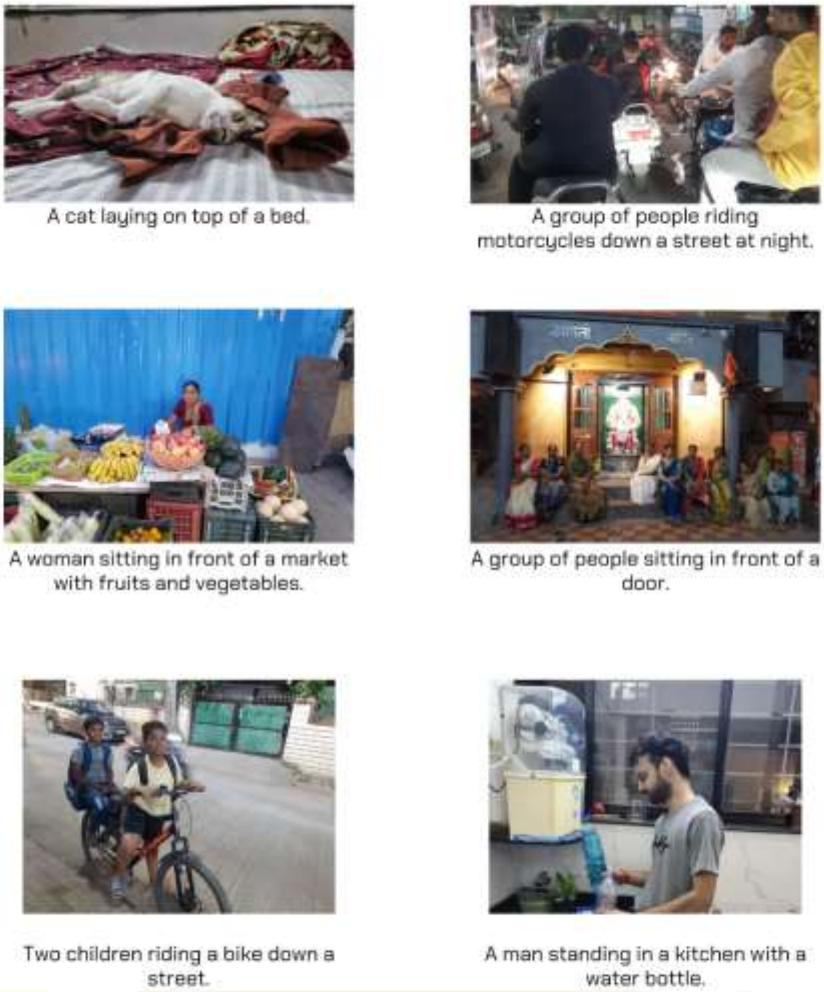
The integration of optical character recognition (OCR) technology enabled the system to extract and read text embedded within images in real-time. Utilizing Python-tesseract, the system accurately detected and extracted text from various sources, including printed materials and digital displays. Visually impaired individuals could access printed information effortlessly, further facilitating independent navigation and information retrieval.

3. Color Detection and Identification:

The color detection module, employing computer vision techniques, provided users with information about the colors present in their environment. By capturing and analyzing images using OpenCV [2], the system accurately detected and identified colors, conveying the information audibly to the user. This feature enabled visually impaired individuals to perceive and appreciate the visual characteristics of their surroundings, enhancing their sensory experience and environmental awareness.

Overall, the Caption Speak project demonstrated significant advancements in assistive technology for visually impaired individuals, providing a transformative solution to address the challenges they face in navigating their daily lives. By delivering real-time descriptive feedback through a user-friendly interface, the system empowers users to interact with their environment more effectively and independently. Through continuous refinement and user feedback, the Caption Speak project aims to further enhance its functionality and accessibility, ultimately improving the quality of life for visually impaired individuals worldwide.

International Research Journal
Research Through Innovation



A cat laying on top of a bed.

A group of people riding motorcycles down a street at night.

A woman sitting in front of a market with fruits and vegetables.

A group of people sitting in front of a door.

Two children riding a bike down a street.

A man standing in a kitchen with a water bottle.

Fig 4.1: Real-life Examples

V. CONCLUSION

In conclusion, the Caption Speak project tackles the core challenge faced by visually impaired individuals in comprehending visual information within their environment. By devising a real-time assistive system that incorporates image captioning, text reading, and color detection capabilities, the project aims to augment the autonomy and functionality of visually impaired users. Through meticulous experimentation and optimization, the project has effectively implemented a solution that delivers descriptive auditory feedback, allowing users to perceive their surroundings with heightened clarity and comprehension. By bridging the visual divide, the Caption Speak project endeavors to empower visually impaired individuals to navigate their daily lives with self-assurance and independence, thereby promoting inclusivity and accessibility within society.

REFERENCES

[1] Arystanbekov, B., Kuzdeuov, A., Nurgaliyev, S., and et al. Image captioning for the visually impaired and blind: A recipe for low-resource languages. TechRxiv, May 2023.

[2] Bradski, G. The OpenCV Library. Dr. Dobb’s Journal of Software Tools, 2000.

[3] Hu, J. C., Cavicchioli, R., and Capotondi, A. Exploiting multiple sequence lengths in fast end to end training for image captioning. In 2023 IEEE International Conference on Big Data (BigData). IEEE, Dec. 2023.

[4] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and DollÃ¡r, P. Microsoft coco: Common objects in context, 2015.

[5] Masud, U., Saeed, T., Malaikah, H. M., Islam, F. U., and Abbas, G. Smart assistive system for visually impaired people obstruction avoidance through object detection and classification. IEEE Access, 10:13428–13441, 2022.

[6] Rahman, M. M., Islam, M. M., Ahmmed, S., and Khan, S. A. Obstacle and fall detection to guide the visually impaired people with real time monitoring. SN Computer Science, 1(4):219, 2020.

[7] Saoji, S., Singh, R., Eqbal, A., and Vidyapeeth, B. Text recogination and detection from images using pytesseract. Journal of Interdisciplinary Cycle Research, XIII:1674–1679, 08 2021.

[8] Tiwary, T. and Mahapatra, R. P. An accurate generation of image captions for blind people using extended convolutional atom neural network. Multimedia Tools and Applications, 82:3801–3830, 2023.

[9] Tsutsui, S. and Crandall, D. Using artificial tokens to control languages for multilingual image caption generation, 2017.