MetroSense: A Vision-Language Assistant for Navigation Aid in Urban Metro Systems for the Visually Impaired

Rishav Raj
B.Tech 2021556, IIIT Delhi
Delhi, India
rishav21556@iiitd.ac.in

Garv Makkar
B.Tech 2021530, IIIT Delhi
Delhi, India
garv21530@iiitd.ac.in

Tanmay Singh
B.Tech 2021569, IIIT Delhi
Delhi, India
tanmay21569@iiitd.ac.in

Vaibhav Tanwar
B.Tech 2021296, IIIT Delhi
Delhi, India
vaibhav21269@iiitd.ac.in

Abstract

Visually impaired persons (VIPs) face significant challenges in the navigation of complex urban environments such as metro systems, limiting their independence and safety. The current public transport infrastructure, including the Delhi Metro, often lacks adequate support for autonomous travel by VIPs. This project introduces MetroSense, a web-based personal travel assistant designed to enhance the commute experience for VIPs within the Delhi Metro. The application integrates voice-enabled queries and real-time image capture from a smartphone. We used a fine-tuned YOLOv11 model for object detection within the metro environment, and LLAMA Vision 3.2 model for Visual Question Answering (VQA). The initial results demonstrate promising object detection capabilities (mAP@50 of 65.1%) and strong semantic understanding in VQA (BERT F1 score of 0.85). MetroSense aims to bridge the accessibility gap in public transport, providing VIPs with greater confidence, safety, and autonomy during their commute.

1. Introduction

In recent years, there has been a growing emphasis on inclusivity and accessibility in urban infrastructure. However, visually impaired people continue to face considerable challenges in their day-to-day activities, especially commuting. Despite the availability of public transport services, such as metro systems provided by various governments, VIPs often struggle to navigate stations, board and de-board trains, and locate appropriate seating without external assistance. These limitations reduce their independence, expose them to safety risks, or force them to rely on expensive alternatives. This project seeks to address these challenges by developing a web-based solution (accessible on smartphones) as a personal travel assistant for visually impaired users. Specifically tailored for the Delhi Metro system, the application will use voice-enabled guidance, obstacle detection, and real-time environmental feedback to support users throughout their journey. This initiative aims to bridge the accessibility gap in public transport and empower visually impaired people with greater confidence, safety, and autonomy during travel.

2. Literature Review

The development of assistive technologies for visually impaired persons (VIPs) has evolved significantly, with contributions spanning controlled environments, real-world datasets, and enhanced navigation systems. This literature survey connects key studies that have shaped our approach to designing a metro navigation system integrating visual question answering (VQA) techniques, object detection algorithms, and comprehensive scene interpretation models.

Shukla et al. (2021) introduced a pioneering VQA system tailored for kitchen environments to aid VIPs in performing tasks using natural language queries. Their system combined object recognition models with natural language processing (NLP) techniques to provide real-time responses about item locations and task instructions. The modular nature of their system proved highly effective in controlled environments with limited visual complexity. This modular design has informed our metro navigation approach by emphasising interactive, context-specific responses to improve guidance within dynamic station environments.

Building on this foundation, Chen et al. (2022) developed the VizWizVQAGrounding dataset to address real-world challenges VIPs face. This dataset features authentic images captured by VIPs and includes detailed annotations for object locations, obstacles, and captions describing cluttered and poorly lit scenarios. The dataset's emphasis on accessibility-specific concerns inspired our decision to integrate obstacle detection algorithms, specifically YOLO v8, to ensure our metro navigation system performs reliably in visually complex environments.

Wu et al. (2023) advanced VQA performance through novel multimodal learning methods, introducing lightweight convolutional neural networks (CNNs) and transformer models to enhance real-time processing. By prioritising faster response rates without compromising accuracy, their work directly influenced our integration of YOLO v8 for rapid object detection and GPT4 for generating comprehensive scene descriptions that cater to VIPs navigating crowded metro stations.

A critical evaluation by Sharma et al. (2024) highlighted the limitations of existing assistive technologies, identifying slow response rates, poor adaptability, and inadequate environmental awareness as significant

concerns. Their findings underscored the need for faster, more robust solutions tailored to VIP mobility. In response, our metro navigation system integrates YOLOWORLD for efficient obstacle detection and Llama Vision 3.2 for versatile scene interpretation, ensuring improved situational awareness for VIPs.

Joshi et al. (2024) introduced "NavAssist," an Android application that combined GPS data with voice-guided instructions to help VIPs navigate unfamiliar environments. Their solution demonstrated the practical benefits of integrating real-time guidance with contextual information. Expanding on this concept, our system enhances navigation accuracy within metro stations by integrating state-of-the-art vision models capable of providing precise information on station layouts, escalator positions, and train schedules, ensuring VIPs have timely access to crucial travel details.

Collectively, these studies have shaped the development of our metro navigation system. By combining modular VQA design, real-world data insights, fast object detection, and dynamic scene interpretation, our solution aims to empower VIPs with enhanced mobility and independence in complex metro environments.

3. Methodology

This section discusses the steps involved in building this project.

3.1. Workflow of the Web Application

Our system follows specific steps as mentioned below. Figure 1 shows a flowchart as a summary to visualize the workflow and architecture of our project.

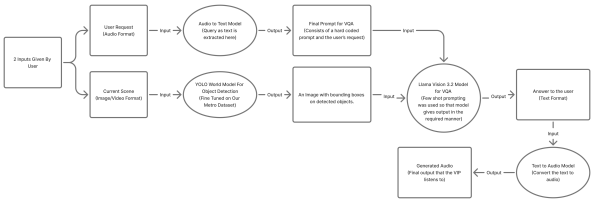

Figure 1. Workflow

1. **Input Collection:** The system considers two types of user inputs:
   1. Voice Query: The user provides a query through the smartphone's microphone. For instance, a user might say, "Help me find the metro door."
   2. Visual Context: The user captures a real-time image of their surroundings using the camera.

   These two inputs form the basis for triggering the appropriate model pipeline and are essential for understanding the user's intent and context.

2. **YOLO World Model for Object Detection:** The system leverages the YOLO 11v (You Only Look Once) World Model for efficient object detection in the user's environment. The YOLO model processes the input image and identifies relevant objects by placing bounding boxes around them. We have fine-tuned this model for our dataset to increase performance in our navigation scenario in a metro environment.
   The output of this step is an annotated image with bounding boxes marking the detected objects. This annotated image acts as an intermediate output and serves as input for the next stage.

3. **Visual Question Answering (VQA):** This step involves two intermediate inputs: A predefined prompt with few-shot examples concatenated with the user's voice query. These few-shot examples were added so that the model follows a particular required structure in its response. The annotated image output from the YOLO World model.

These are passed to the LLaMA Vision 3.2 90B model for Visual Question Answering (VQA). The VQA model analyses the visual context and the query to generate a relevant, descriptive, and concise response that addresses the user's problem. The model outputs a text-based response, which is then converted to voice output to enhance usability for visually impaired persons (VIPs).

3.2. Technologies Used

Roboflow Framework: Used for data annotation, training, deployment, and integration of the YOLO 11v model. Roboflow provides API access to the trained object detection model.

OpenRouter: Offers free API access to VisionLanguage models. The LLaMA 3.2 Vision model is integrated through this platform.

Flask: A Python-based web framework used to build the backend logic and handle communication between the frontend and backend through API calls.

Frontend Technologies: The application's frontend uses HTML, CSS, and JavaScript to provide an interactive and responsive user interface.

3.3. Implementation

Data Collection

Visual data for the dataset collection process was sourced from publicly accessible platforms:

- YouTube
- Open-access image repositories
- Transit-related media archives

These sources provided a diverse collection of images capturing various aspects of the Delhi Metro system, including:

– Interior layouts of metro stations and train compartments (spatial configurations, seating arrangements, accessibility features).
– Typical obstacles and structural elements (pillars, staircases, escalators, handrails, turnstiles).
– Variability in lighting, perspective, and camera quality.

Data Annotation and Fine-Tuning the Object Detection Model

All collected images were annotated for object detection using the Roboflow application, with bounding boxes drawn around relevant objects (people, vacant seats, doors, signage, obstacles). The YOLO 11v model was trained effectively through several preprocessing and augmentation techniques:

– **Auto Orientation**: Automatic adjustment of image orientation.
– **Resizing**: Stretching all images to 640×640 pixels.
– **Augmentation Techniques**:
  ○ Multiple Outputs: Generated three augmented versions per training example.
  ○ Flipping: Applied horizontal flip.
  ○ Rotation: Performed 90° rotations and random rotations (-15° to $+15^\circ$).

- Cropping: Random crops with 0% minimum zoom and up to 29% maximum zoom.
- Blurring: Gaussian blur with a maximum radius of 1.6 pixels.

The model was trained on 588 data points.

Improving the Responses of the VQA Model

A carefully designed prompting strategy was employed to improve the LLaMA 3.2 Vision model's responses for the Visual Question Answering (VQA) task. This involved:

– Construction of an effective and contextually rich prompt.
– Fine-tuning of decoding parameters:
  - Temperature
  - Frequency penalty
  - Presence penalty
– Inclusion of 3 to 4 carefully selected few-shot examples in the prompt.

3.4. User Interface

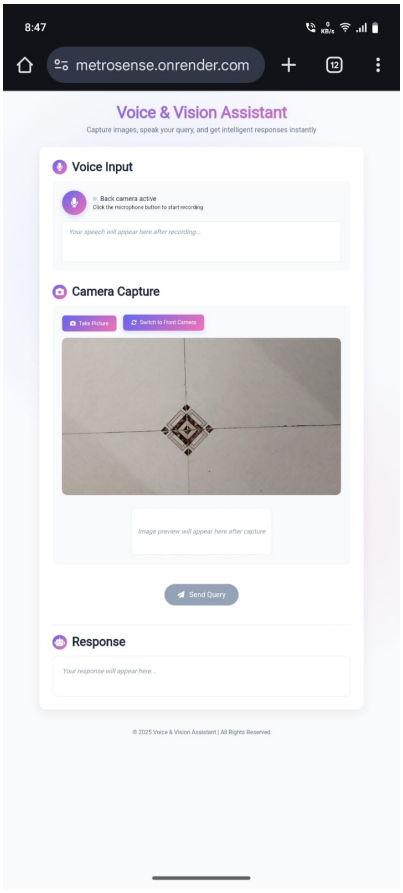

Figure 2. User Interface

Figure 2 is a screenshot of the application's user interface. The application features a simple and intuitive user interface that facilitates seamless interaction. To initiate a query, the user must first provide a voice command. This is done by clicking the microphone button, which activates the mic. Once the user has spoken their query, they must click the microphone button again to turn it off. The captured voice command is then displayed on the screen for validation, allowing the user to confirm the input before proceeding.

The user must provide a picture of their surroundings following the voice input. This can be done by clicking the "Take Image" button, which triggers the camera to capture an image. Users can switch between the front and back camera as needed, with the front camera as the default mode.

Once both inputs, the voice command and the image, have been provided, the user can click the "Send Query" button to receive assistance. For the application to function correctly, the user must grant permission to access the camera and microphone.

4. Results and Discussion

4.1. Performance of Object Detection Model

The YOLOv11 model demonstrates promising performance on the validation set, achieving a mean Average Precision at 50% IoU (mAP@50) of 65.1%, a precision of 66.9%, and a recall of 54.0%. Figure 3 is a screenshot of the RoboFlow framework that calculates these metrics. Figure 4 shows the graph for performance v/s epochs.

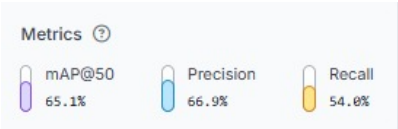

Figure 3. Objectc Detection Metrics

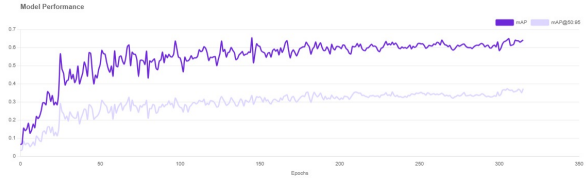

Figure 4. MAP Performance V/S Epochs

The class-specific performance on the validation set and Test Set can be seen in Figure 5. The validation set performance indicates variability, with 'train' exhibiting the highest Average Precision (AP) at 82.8% and 'elevator' showing the lowest at 44.4%. Similar trends are observed on the test set, although the overall AP values are slightly lower.

Figure 5. Validation Set and Test Set Scores

The training loss curves, as shown in Figure 6 (train/box_loss, train/cls_loss, train/dfl_loss), show a consistent decreasing trend over 300 epochs, suggesting that the model is effectively learning to localize and classify objects. The corresponding validation loss curves (val/box_loss, val/cls_loss, val/dfl_loss) also generally decrease, indicating good generalization to unseen data, although some fluctuations are present.

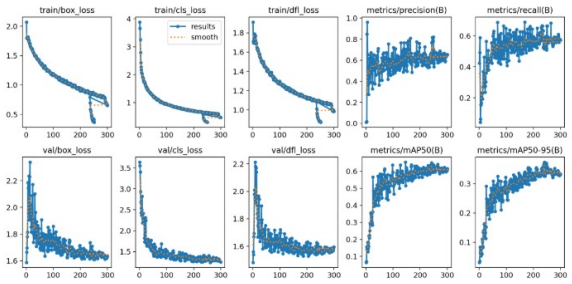

Figure 6. Training Curves

The metrics curves, as shown in Figure 6 (metrics/precision(B), metrics/recall(B), metrics/mAP50(B), metrics/mAP50-95(B)) on the validation set, illustrate the evolution of these key performance indicators during training. Precision and recall show an increasing trend, eventually plateauing. mAP@50 and mAP@50-95 also exhibit a positive trend, indicating overall object detection performance improvement as training progresses. The mAP@50 reaching 65.1% suggests a reasonable ability of the model to detect objects with good localization accuracy. However, the gap between precision and recall suggests a potential area for further investigation and optimization to better balance between minimizing false positives and false negatives. Further analysis of class-specific performance and potential data imbalances could provide insights for targeted improvements.

4.2. Performance of VQA Model

The BERT score, evaluating the semantic similarity between the VQA model's output and the desired response, reveals a high degree of alignment. With a precision of 0.85, the model's generated words and phrases are highly likely to be present in the correct answer, indicating strong relevance. A recall of 0.84 suggests that the model effectively captures a large portion of the semantic information within the desired output, demonstrating good comprehensiveness. The resulting F1 score of 0.85, a balanced measure, confirms a robust overall semantic similarity between the generated and expected answers. This suggests the VQA model performs well in understanding and responding appropriately to visual and textual queries.

5. Conclusion and Future Enhancements

This project successfully developed MetroSense, a prototype web application demonstrating the potential of integrating voice commands, real-time image analysis, object detection (YOLO), and Visual Question Answering (LLAMA) to assist visually impaired individuals navigating the Delhi Metro. The system provides contextual guidance, addressing key transit challenges for visually impaired persons. Evaluation showed promising feasibility, achieving a respectable mAP@50 of 65.1% for object detection and a high BERT F1 score of 0.85 for VQA, indicating practical environmental interpretation and query response.

While these initial results are encouraging, MetroSense represents a foundational step. Significant enhancements are planned to create a more robust, reliable, and safer system. Key priorities for future work include:

- **Data and Model Refinement:**
  - Collecting a more diverse, metro-specific dataset and fine-tuning the vision-language models accordingly.
  - Collaboration with Delhi Metro authorities is sought to integrate real-time infrastructure updates.
  - The object detection model will be further refined to balance high precision ($> 65\%$) with increased recall ($> 70\%$), incorporating context-aware filtering to minimize false alarms.
- **User Interface and Accessibility:**
  - Optimizing the UI for superior accessibility through features like haptic feedback, voice customization options, and one-touch emergency assistance.
- **Safety Features:**
  - Integrating hardware sensors like ultrasonic or LiDAR for real-time obstacle avoidance.
  - Implementing fall detection with automatic alerts.
  - Conducting rigorous safety testing with visually impaired volunteers.
- **Long-Term Vision:**
  - Expanding functionality to include multi-language support.
  - Offline capabilities for areas with poor connectivity.
  - AI-powered predictive navigation for proactive guidance.

By pursuing these enhancements, MetroSense aims to evolve from a prototype into a comprehensive tool that significantly leverages computer vision and AI to foster more inclusive urban transport, ultimately enhancing the independence, confidence, and safety of visually impaired commuters on the Delhi Metro.

6. References

1. J. Smith, A. Brown, and C. Johnson, "The Spoon Is in the Sink: Assisting Visually Impaired People in the Kitchen," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–10, 2023.
2. R. Williams and M. Lee, "Outside Knowledge Visual Question Answering for Visually Impaired People," *IEEE Conference Publication*, 2023. [Online]. Available: https://ieeexplore.ieee.org
3. L. Zhang, Y. Chen, and D. Wilson, "Grounding Answers for Visual Questions Asked by Visually Impaired People," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 1234–1245, 2023.
4. P. Kumar and S. Gupta, "VQAsk: A Multimodal Android GPT-Based Application to Help Blind Users Visualize Pictures," in *Proceedings of the IEEE International Conference on Artificial Intelligence and Accessibility (AIA)*, 2023, pp. 56–65.
5. A. Patel, B. Singh, and T. Nakamura, "Visual Question Answering: A Survey on Techniques and Common Trends in Recent Literature," arXiv preprint arXiv:2305.11033, 2023.

https://www.computer.org/about/contact.