Sub-team 26.3 Report: Daniel & James

Part 1 – Planning

Identifying the Problem/Improvement
A primary challenge for visually impaired users is obtaining clear, real-time information about their surroundings. Our goal for this application is to provide rich, natural-language descriptions of each scene, specifying details such as color, distance, or an object's context. In our current web app, users must manually upload images for processing, which can be inconvenient or impractical for someone who cannot easily see the screen or locate the necessary interface elements.

To address these limitations, we are introducing two major capabilities:

1. **Image-to-Description (GPT-4 Vision):** Instead of basic labels like "Car," the system will produce more context-aware observations, e.g., "There is a red Toyota parked approximately three meters to your left."
2. **Speech Interface (Speech-to-Text and Text-to-Speech):** Users can speak queries (e.g., "Is it safe to cross the street?") and receive responses without manually uploading images or typing anything.

These capabilities leverage AI for improved contextual accuracy and a more hands-free, accessible experience.

Criteria, Research, and Selection
When choosing AI tools, we focused on accuracy, latency, and how easily they integrate with our Node.js/React/Next.js stack. For image-to-text, GPT-4 Vision was more suitable than combining YOLO with a separate language model because GPT-4 Vision can deliver detailed scene descriptions without requiring us to train or maintain multiple models. We opted for Whisper (OpenAI) for speech recognition due to its robust performance across varied accents and noise levels. For text-to-speech, OpenAI TTS delivers natural-sounding voices, offers easy integration, and allows users to pick among multiple voice options.

Regarding ethics and feasibility, using pre-trained models means there is no need for large datasets or extensive training, which simplifies implementation and reduces privacy concerns. We still ensure secure handling of user data (images and audio) to protect personal information.

Why It Aligns With Project Goals
Our project aims to empower visually impaired individuals by helping them navigate their environment with greater ease. Detailed descriptions of what the camera sees and the ability to issue voice commands both significantly reduce the friction of traditional interfaces—like having to type queries or manually upload files. These features directly support the objective of providing a smoother, more autonomous user experience.

Part 2 – Impact Analysis

Improvement Over Baseline
Previously, our application required users to manually upload images to receive descriptions, which could be inconvenient for individuals with visual impairments. Without AI-driven enhancements, users would have to rely on external tools or manually type queries to get more information about their surroundings. By integrating GPT-4 Vision, the system now provides detailed, natural-language descriptions that include color, approximate distance, and contextual details. Additionally, the speech-to-text and text-to-speech features eliminate the need for manual input or visual navigation of the UI. Now, users can simply ask, “What’s in front of me?” and receive an immediate spoken response, making the experience more intuitive and accessible.

Testing and Validation
We will verify the accuracy of our system by ensuring that the speech-to-text component correctly transcribes user input, that GPT-4 Vision accurately describes the environment, and that the text-to-speech output correctly conveys the generated response. These tests will be conducted in controlled environments, allowing us to systematically evaluate the reliability of each component.

Trade-Offs and Considerations
Using AI services may introduce additional latency and costs if multiple requests occur simultaneously. To address latency, we will optimize image resolution and use caching where appropriate. We also recognize that calls to GPT-4 Vision, Whisper, and OpenAI TTS incur API costs, so we need to monitor usage carefully. Despite these trade-offs, the improvements to user experience, safety, and autonomy for visually impaired individuals make this approach well worth pursuing.