Real-Time Visual Understanding Assistance for the Blind

AI Problem Statement:

The core problem is that visually impaired individuals struggle to understand and navigate their surroundings because current assistive technologies often provide only static descriptions or limited contextual information. For example, while there are tools that can recognise certain objects or read out text in an image, they do not support meaningful back-and-forth conversation about what the user is "seeing" through a camera feed. The user might know there is a table in front of them but not know what's on the table, what's next to the coffee cup, or whether the seat at the table is free. Similarly, in complex public spaces like airports or grocery stores, simply announcing an object's existence isn't enough—users may need to ask additional questions about brands, signage, or how to find specific items or facilities.

Those few applications that offer some of the advanced features are extremely expensive thus inaccessible to the larger general population.

The problem statement calls for developing a solution that goes beyond static recognition. Instead of just listing objects, this system should let the visually impaired user engage in a natural, spoken dialogue with an AI assistant that "sees" for them. The assistant will not only identify objects and read signs but also answer context-specific follow-up questions. For instance, after stating that there is a coffee cup on the table, the user might ask, "What's next to the coffee cup?" If it's a plate of cookies, the assistant should respond accordingly. The user might then ask, "Are there any available seats nearby?" and get a direct, helpful answer. In a grocery store, a user might point their camera at a cereal aisle and ask, "What brand is that box with the red label?" or "Is there a discount sticker on that product?"

To achieve this, the proposed solution involves combining multiple cutting-edge AI technologies: computer vision to detect objects and read text, natural language processing (NLP) to understand user questions and maintain context across multiple queries, speech recognition and text-to-speech modules for a seamless voice interface, and possibly machine learning models that learn from user feedback over time to improve accuracy and relevance of responses.

In short, the problem is about creating a real-time, context-aware, conversational system that helps visually impaired users gain a richer, more interactive understanding of their environment—one that feels like having a guide who can see, describe, and discuss the world around them.

The goal is to create a unified solution that:

- **Provides real-time scene interpretation:** Capturing live camera feeds and returning immediate, descriptive audio feedback about what's visible (e.g., identifying objects, reading signage, detecting free spaces to sit).
- **Enables context-sensitive Q&A dialogues:** Users can ask follow-up questions, both general (e.g., "What's on the table?") and context-specific (e.g., "What is next to the coffee cup?"), and receive detailed responses that consider previously identified objects, relative positions, and attributes (such as color, brand, and functionality).
- **Learns and improves over time:** By incorporating user feedback, the system becomes more accurate, fine-tuning object recognition, language understanding, and the precision of answers in complex, real-world environments.

By focusing on rapid prototyping with advanced AI tools, participants can deliver a proof-of-concept that showcases the feasibility and transformative potential of such a solution within a 10-day hackathon timeframe.

Plan of Action (10-Day Hackathon):

Day 1-2: Requirements & Data Setup

Define user stories and core functionalities (e.g., object detection, multi-turn Q&A, speech-to-text, text-to-speech).

Gather a small dataset of images with annotated objects and scene descriptions. Include everyday objects, public space elements (signs, tables, brand logos), and various lighting conditions.

Day 3-4: Model Integration Baseline

Implement a pre-trained object detection model (e.g., YOLOv5 or EfficientDet) to identify objects in real-time camera feeds.

Integrate a robust ASR (Automatic Speech Recognition) component (e.g., Whisper or wav2vec2) to convert voiced queries into text.

Use a text-to-speech engine (e.g., Amazon Polly, Google Cloud TTS) for producing clear audio responses.

Day 5: Image Captioning & Initial Q&A

Add a pre-trained image captioning model (e.g., a ViT + Transformer-based captioner) to generate initial scene descriptions.

Implement a baseline NLP pipeline for Q&A. Start with a rule-based approach or a lightweight transformer model fine-tuned on relevant caption-QA pairs. This enables initial multi-turn dialogues.

Day 6: Context-Aware Reasoning & Dialog State Tracking

Introduce dialog state tracking to maintain context over multiple queries. For example, when the user asks, “What’s next to the coffee cup?”, the system should refer back to the previously identified objects.

Fine-tune or use a small LLM (Large Language Model) for improved conversational coherence and to handle variations in user queries, ensuring the system understands relationships (e.g., relative positions) and attributes (color, brand, size).

Day 7: Rapid Feedback Loop & Personalisation

Implement a simple user feedback mechanism (e.g., voice command “That’s correct” or “Not quite”) to record whether the user is satisfied.

Adjust prediction thresholds or prompt conditioning to improve accuracy based on feedback.

Begin testing in low-contrast and noisy environments to ensure robustness.

Day 8: Public Space Adaptations & Brand/Sign Recognition

Add a text recognition (OCR) component for reading signs, labels, or product branding in public spaces (e.g., Tesseract or Vision OCR APIs).

Integrate a few known brand logos to test brand-specific Q&A capabilities ("What brand is on that cereal box?").

Test the system in simulated scenarios of museums and airports (e.g., guiding users to an exit sign or identifying free tables).

Day 9: Latency Optimisation & UI/UX Polishing

Optimise inference speed by model pruning, quantisation, or batching requests to ensure near-real-time performance.

Refine the user interface, ensuring intuitive voice commands and seamless feedback loops.

Improve error handling—if the system can't identify an object or answer a question, it should gracefully prompt the user for clarification.

Day 10: Final Integration, Testing, and Presentation

Conduct end-to-end tests under various conditions (poor lighting, crowded scenes, background noise).

Finalise a demo scenario: the user entering a cafe, asking about available seats, identifying items on a table, and following up with questions about brands or colours.

Prepare a concise, compelling presentation highlighting the prototype's capabilities, the technology stack, and future potential.

Technological Path:

Computer Vision:

Object detection using state-of-the-art CNN or transformer-based models (e.g., YOLO, DETR).

OCR for text in the environment, enabling brand recognition and reading signs.

Natural Language Understanding:

Automatic Speech Recognition (ASR) for capturing user's spoken queries.

NLP-based dialogue systems using transformer-based language models (e.g., a fine-tuned BERT, T5, or a small LLM) for Q&A and maintaining conversational context.

Text-to-Speech and Audio Feedback:

High-quality TTS solutions (cloud-based or on-device) for immediate, clear spoken responses.

Continuous Learning and Feedback:

Simple reinforcement signals from user corrections to improve future responses.

Metadata logging (object recognised, question asked, response given) to refine models post-hackathon.

This 10-day rapid prototyping challenge demonstrates the feasibility and value of an AI-driven, context-aware, visually assistive system. By integrating cutting-edge models with iterative user feedback, participants will showcase a working proof-of-concept that can serve as the foundation for a fully-realised product, ultimately empowering visually impaired individuals to navigate and understand their environments with greater independence and confidence.

Rules of Engagement:

Intellectual Property Ownership:
All solutions and associated work products created during the hackathon directly address Mythicka's flagship product needs(already conceptualised). While students may be recognised as co-authors of any Intellectual Property (IP) developed, all IP rights—including copyrights, patents, trademarks, and any other IP that may arise—will remain with Mythicka. Participation in the hackathon constitutes an understanding and acceptance of this arrangement.

Handover of Deliverables:
Upon completion of the hackathon, students will provide Mythicka with all developed materials, including but not limited to source code, documentation, design assets, and trained model files. These deliverables become the property of Mythicka for further development, commercialisation, or any other purpose Mythicka sees fit.

No Formation of External Entities:
Students are not invited, required, or encouraged to form any separate entity—such as a startup or company—based on their hackathon solutions. The objective is to solve Mythicka's defined problem statements, and any external commercialisation or separate venture creation is not authorised.

Involvement Beyond the Hackathon:
If Mythicka wishes to further develop a promising solution into a Minimum Viable Product (MVP) or beyond, it may invite participating students to continue their involvement. Such an invitation does not imply or grant any equity, shares, or ownership stake in Mythicka. Any terms of engagement, whether paid or otherwise, will be determined at Mythicka's sole discretion, with no entitlement inferred for the students.

Recognition and Credit:
While Mythicka retains ownership, Mythicka may publicly acknowledge and credit students for their contributions, such as through mentions in official communications, press releases, or product documentation. Such recognition will not alter the agreed-upon IP or ownership rights.

By agreeing to champion this problem statement, you agree to the above terms and conditions of engagement.