BABEȘ-BOLYAI UNIVERSITY, CLUJ-NAPOCA, ROMÂNIA

Faculty of Mathematics and Computer Science

Real-Time Scene-Aware Assistant

for Blind People using TinyVLM

MIRPR Project Report 2025–2026

Team Members:

Moglan Călin
Puscas Raul

Cluj-Napoca, 2025–2026

Abstract

This project proposes the development of a **Real-Time Scene-Aware Assistant** for blind or visually impaired individuals using **TinyVLM** — a compact Vision-Language Model optimized for edge deployment. The system performs continuous real-time analysis of the environment, generating detailed, context-aware natural language descriptions and answers to user queries. The project emphasizes **autonomy, safety, and accessibility** through efficient AI models deployable on mobile or wearable devices. Experiments will focus on recognition accuracy, natural language quality, and real-time performance on low-power devices.

Contents

1. Introduction

1.1 What? Why? How?

The **Real-Time Scene-Aware Assistant for Blind People** aims to provide intelligent visual assistance using the **TinyVLM** architecture.

What is the problem?

Blind or visually impaired individuals face significant challenges in understanding and interacting with their surroundings. Tasks such as identifying obstacles, recognizing people, or understanding environmental context require external assistance.

Why is it important?

This project directly impacts accessibility and inclusion by enhancing independence and safety for visually impaired users. Moreover, it demonstrates the power of **efficient AI on edge devices**, proving that advanced multimodal reasoning can operate within limited hardware constraints.

How do we address it?

We design an application that continuously processes camera input, interprets scenes through TinyVLM, and provides real-time natural language descriptions or spoken responses.

Key Base Functionalities

- Real-time environmental analysis.
- Context-aware natural language scene descriptions.
- Dynamic object and movement interpretation.
- Responsive question-answering system.
- Facial recognition for familiar people.

1.2 Paper Structure and Contributions

The report presents:

- A comprehensive analysis of challenges faced by visually impaired individuals.
- A real-time TinyVLM-based algorithm for scene understanding.
- A deployable prototype on mobile and wearable devices.

2. Scientific Problem

2.1 Problem Definition

Blind or visually impaired people lack real-time access to visual information essential for safe navigation and interaction. Existing systems often rely on bulky equipment or cloud processing, leading to latency and privacy concerns.

Input: Continuous video feed from a wearable or mobile camera, and optional voice queries. **Output:** Natural language descriptions, spoken alerts, and answers to user questions.

Performance Criteria

- Object/action recognition accuracy (mAP, precision, recall)
- Description quality (BLEU, METEOR, CIDEr, SPICE)
- Latency and real-time response
- Efficiency on low-power hardware
- Human evaluation — perceived usefulness and satisfaction

3. Related Work

The following research highlights the advancements that inspired this project:

1. **Vision-Language Models for Edge Networks:** Lightweight architectures optimized for mobile and IoT deployment.
2. **Navigation with VLM Framework (NavVLM):** Integrates VLM reasoning into robot navigation.
3. **MoViNets:** Efficient 3D CNNs for real-time video recognition.
4. **TinyVLM:** Compact VLM (0.6B parameters) enabling real-time CPU inference.
5. **Vision Transformer (ViT):** Transformer-based visual representation learning foundational to VLMs.

4. Application Study Case

4.1 Description and Functionalities

The application runs on mobile or wearable devices, providing real-time scene interpretation and natural language feedback.

- Visual context description and event detection.
- Voice-activated question answering.
- Facial recognition for known individuals.
- Real-time feedback via text-to-speech.

4.2 Example Use Cases

- “Describe my surroundings.”
- “Is there a person in front of me?”
- “Who is nearby?”

5. Conclusion and Future Work

This project demonstrates that small-scale Vision-Language Models like **TinyVLM** can deliver real-time, socially impactful AI solutions on low-power devices.

Future work includes:

- Integrating multimodal sensors (audio, GPS).
- Enhancing dialogue memory for natural interaction.
- Extending multi-language support.