DATA SCIENCE & AI LAB (BSCSS3001)

MILESTONE - 3: MODEL ARCHITECTURE

GROUP NO. 2

PRASHASTI SARRAF **(21f1001153)**
TANUJA NAIR **(21f1000660)**
BALASURYA K **(22f3002744)**
KARAN PATIL **(22f2001061)**
JIVRAJ SINGH SHEKHAWAT **(22f3002542)**

IITM BS Degree Program
Indian Institute of Technology,
Madras, Chennai,
Tamil Nadu, India, 600036

Vision Assist: Real-Time Navigation Support for the Visually Impaired

Model Architecture

1. Abstract

This milestone focuses on identifying and justifying the model architecture for real-time object perception, depth estimation and audio feedback to assist visually impaired individuals.

The system integrates **YOLO-World** for open-vocabulary object detection, **DeepSORT** for motion tracking, **MiDaS** for depth estimation, and **Coqui TTS (Tacotron2 + HiFi-GAN)** for speech-based output. Together, these components form an intelligent assistive pipeline that detects, tracks, measures, and narrates environmental context to the user for safe and autonomous navigation.

The data used for training and fine-tuning the models is sourced from the **COCO dataset**, which provides labeled everyday objects, and **frames extracted from YouTube videos**, which capture real-world, dynamic scenarios. This combination ensures that the system can generalize well to diverse and unpredictable environments encountered in daily life.

2. System Overview

Component	Purpose	Selected Model	Output
Object Detection	Detect and label surrounding objects	YOLO-World	Bounding boxes, object classes and confidence
Object Tracking	Maintain consistent object identity across frames	DeepSORT	Unique object IDs and movement direction
Depth Estimation	Estimate relative distance of objects from the user	MiDaS	Depth map and distance approximation
Voice Output	Provide natural audio guidance	Coqui TTS (Tacotron2 + HiFi-GAN)	Spoken feedback

3. Model Architecture Overview

YOLO-World

- **Type:** Open-vocabulary object detection model (extension of YOLO architecture).
- **Key Idea:** Leverages CLIP embeddings to recognize objects beyond the training dataset.
- **Output:** Bounding boxes + confidence scores for both known and unseen object categories.

DeepSORT

- **Type:** Object tracking algorithm that uses motion (Kalman Filter) + appearance (CNN embedding).
- **Key Idea:** Assigns unique IDs to each detected object and tracks them across frames.
- **Output:** Continuous trajectories for moving objects.

MiDaS

- **Type:** Monocular depth estimation network.
- **Key Idea:** Predicts pixel-wise depth maps from a single RGB image using encoder–decoder CNN.
- **Output:** Relative distance values representing scene geometry.

Coqui TTS (Tacotron2 + HiFi-GAN)

- **Type:** Neural Text-to-Speech (TTS) model.
- **Key Idea:** Converts scene-text descriptions (e.g., “Person two meters ahead”) into natural audio speech.
- **Architecture:**
    - **Tacotron2:** Seq2seq acoustic model generating mel-spectrograms.
    - **HiFi-GAN:** Neural vocoder converting spectrograms to waveform.
- **Output:** Real-time, human-like speech cues guiding user navigation.

4. Justification of Chosen Models

Why YOLO-World?

YOLO-World enables the system to detect a **wide range of real-world objects** using natural language prompts. This helps a visually impaired person identify everyday items like *“wheelchair,” “crosswalk,”* or *“bench”* without needing to retrain the model.
It can also **focus on a specific object** — for instance, if a user wants to find their *“chair”* or *“bag”* in a room, the system can detect only that, reducing distractions. Its **real-time speed** ensures instant feedback, and since it runs efficiently on mobile or wearable devices, users can rely on it anywhere.
Real-time suitability: High – optimized lightweight variant suitable for edge/mobile devices.

Why DeepSORT?

DeepSORT helps the system **track moving objects smoothly** across frames. This ensures stable and consistent detection — for example, it can alert the user that a *“car is approaching from the left”* or a *“person is crossing ahead.”*
By maintaining motion consistency and preventing flickering, it provides **reliable movement cues** that enhance confidence and safety while walking outdoors.
Real-time suitability: High – low computational overhead enables real-time frame-by-frame tracking.

Why MiDaS?

MiDaS adds **depth awareness** using just a regular camera. It estimates how close or far an object is and gives **distance-based alerts** such as *“obstacle two meters ahead”* or *“wall one meter away.”*
This helps users **judge space and distance**, allowing them to navigate safely without colliding with nearby objects.
Real-time suitability: Moderate-High – single RGB input, fast inference (~20–25 FPS on GPU, slower on CPU).

Why Coqui TTS?

Coqui TTS is an advanced open-source text-to-speech framework derived from Mozilla’s TTS project. It provides state-of-the-art speech synthesis capabilities using deep learning architectures such as **Tacotron 2**, **VITS**, and **Glow-TTS**, combined with high-quality **neural vocoders** like **HiFi-GAN** and **MelGAN**.
It generates clear, natural, and offline-capable voice feedback for detected scene descriptions, and enhances accessibility and usability in real-world assistive settings. Its modular design allows rapid experimentation and deployment, making it ideal for both cloud-based and **on-device (edge)** applications.
Real-time suitability: High – optimized for fast TTS inference on CPU/GPU for edge deployment.

5. Model Comparison: Advantages and Limitations

Task	Selected Model	Alternatives	Advantages of Selected	Limitations / Trade-offs
Object Detection	YOLO-World	DETR / ReDETR, SSD, Faster R-CNN, EfficientDet	• Open-vocabulary detection via CLIP embeddings → detects unseen objects • Real-time speed suitable for edge/mobile devices • High mAP and robust bounding boxes	• Slightly heavier than YOLOv8 • Requires GPU for best performance • Prompt sensitivity
Object Tracking	DeepSORT	SORT, ByteTrack, FairMOT	• Combines motion (Kalman filter) + appearance embeddings for stable tracking • Maintains consistent	• Can lose track under long-term occlusion • Dependent on detection quality

			object IDs even during short occlusions • Lightweight and widely used	
Depth Estimation	MiDaS	Monodepth2, DenseDepth, DepthFormer	• Accurate pixel-wise depth from monocular RGB • Robust to diverse environments (indoor/outdoor) • Lightweight for near real-time inference	• Relative depth only (needs scaling for absolute distance) • Some loss of detail for very small or distant objects
Text-to-Speech	Coqui TTS (Tacotron2 + HiFi-GAN)	Google TTS, Amazon Polly, VITS, Kitten TTS	• Open-source and offline capable → no internet dependency • High-quality, natural-sounding speech • Supports real-time inference on CPU/GPU • Flexible voice customization and multilingual support	• Initial model download size (~1–2 GB) • Slightly higher latency on CPU compared to cloud TTS APIs

Key Takeaways

- **YOLO-World** over DETR/SSD/Faster R-CNN/EfficientDet:
  Chosen for **open-vocabulary support** and **real-time edge inference**, unlike DETR which is slower and SSD/Faster R-CNN which are limited to predefined classes.

- **DeepSORT** over SORT/ByteTrack/FairMOT:
  Offers **appearance-based re-identification**, reducing ID-switches during motion and short-term occlusions, which basic SORT or ByteTrack cannot handle reliably.

- **MiDaS** over Monodepth2/DenseDepth/DepthFormer:
  Produces **high-quality depth maps across diverse scenes** using a single RGB

frame, making it practical for indoor/outdoor navigation.

- **Coqui TTS** over VITS/Kitten TTS/cloud APIs:
  Provides **offline, real-time, high-fidelity voice synthesis** while being **open-source** and fully customizable — critical for privacy and assistive applications.

6. Integration Workflow

Camera → YOLO-World → DeepSORT → MiDaS → Fusion → Coqui TTS → User

Step 1:
The live camera feed is captured and passed into **YOLO-World** for open-vocabulary object detection.

Step 2:
The detections (bounding boxes and class labels) are then sent to **DeepSORT** for ID assignment and motion tracking.

Step 3:
In parallel, the same frame is processed by **MiDaS** to generate a corresponding depth map.

Step 4:
All outputs — object class, motion trajectory, and distance — are fused together and sent to the **voice output module**, which prepares a concise verbal description (e.g., *“Person moving 2 meters ahead.”*).

Step 5:
The **Coqui TTS module** converts this textual description into natural-sounding speech, delivering real-time auditory guidance to the user.

This integrated pipeline enables a **context-aware assistive system** that can perceive, interpret, and communicate real-world scenes seamlessly in real time.

7. Way Forward

1. Quantize YOLO-World and MiDaS for **mobile deployment** (Jetson, Android).

2. Incorporate **spatialized audio cues** for directional awareness.

3. Add **speech recognition** for user commands (“find door”, “avoid obstacle”).

4. Integrate **Coqui TTS** for ultra-fast real-time speech synthesis in edge devices.

8. Conclusion

The final architecture — **YOLO-World + DeepSORT + MiDaS + Coqui TTS** — provides a robust, multimodal perception and guidance system for visually impaired navigation.

This setup balances accuracy, adaptability, and accessibility, achieving real-time environmental understanding, intuitive speech-based assistance, and enhancing independence for visually impaired users.

Annexure

Github Repository
Drive Data Storage Link (dsai project)