Literature Survey Presentation

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

SUBMITTED BY:
Akash Kushwaha 2021514
Manav Mittal 2021538
Utkarsh Venaik 2021570
Shreyas Kabra 2021563

Introduction

VisionPulse: Revolutionizing Education for Visually Impaired Students

1. **Innovative Solution** designed to support visually impaired students at IIIT Delhi.
2. Leverages the power of **Computer Vision** and **Natural Language Processing** to develop a **Large Multi-Modal (LMM)** system.
3. **Context-aware assistance** for seamless navigation and interaction within the campus.
4. Tailored to enhance **independent mobility**, interaction with **peers**, and engagement with **educational resources**.
5. Bridges the gap in the educational experience by providing **real-time support** for lectures, study materials, and campus activities.
6. Aims to foster an **inclusive environment**, making learning and social integration smoother for visually impaired students.
7. **Campus-specific customization**, designed to adapt to the unique challenges of a university setting.

Research Questions

What technologies are available to and applied to assist the visually impaired in educational settings?

AI, NLP, and Computer Vision are transforming the educational experience for visually impaired individuals:

- **Vision-based Assistive Systems**:
  - Systems like **Seeing AI** and **NavCog** provide real-time auditory feedback and indoor navigation.
- **Natural Language Processing (NLP)**:
  - Tools like **VoiceOver** (iOS) & **TalkBack** (Android) read on-screen content aloud.
  - Advanced models (GPT, BERT) provide context-aware responses, helping visually impaired students access large amounts of text material.
- **Wearable Technologies**:
  - Devices like **OrCam MyEye** and **Envision Glasses** use computer vision to recognize faces, read text, and provide real-time auditory feedback.

These technologies enhance **autonomy, accessibility**, and the **learning experience** for visually impaired individuals.

Research Questions

Where can these technologies be applied most effectively within educational institutions?

Key Areas: Classrooms and Campuses

- **Classroom Settings**:
  - **Challenges**: Accessing visual materials (lecture slides, textbooks).
  - **Solutions**: Tools like **LectureSight** and advanced **NLP models** can transcribe and summarize live lectures, converting spoken content and visual material into accessible audio or text formats for real-time participation.
- **Campus Navigation**:
  - **Challenges**: Navigating complex campus layouts.
  - **Solutions**: Systems like **NavCog** and **Seeing AI** provide **real-time auditory guidance**, using computer vision to direct students through hallways, classrooms, and outdoor areas with step-by-step instructions and visual-to-auditory conversion.

These technologies significantly enhance **learning** and **mobility** for visually impaired students.

Literature Review

Scholarly Databases Used

1. Google Scholar
2. ACM Digital Library
3. IEEE Xplore
4. Application and model documentation from GitHub, HuggingFace, etc.
5. arXiv, etc.

Keyword Searching

- "taxonomy of pain points of visually impaired individuals"
- "artificial intelligence systems for visually impaired people"
- "vision-based assistive technologies for students"
- "multi-modal AI-based systems"
- "facial recognition systems using image matching"
- "voice-command AI systems"
- "NLP for visual assistance"

Etc.

Relevant Studies

Studies were selected based on:

1. Understanding the pain points people with visual disabilities face to understand what our project is solving.
2. Direct relevance to VisionPulse's core engineering functionalities eg: visual question answering (VQA), obstacle detection, STT and TTS systems, contextual information retrieval (RAG), etc.
3. Relevant evaluation metrics and datasets for improving our model performance and compare baselines to.

The paper are divided based on themes

Theme 1:
Understanding pain points of people with visual challenges

Research Paper 1

TITLE: Young Persons with Visual Impairment: Challenges of Participation by Salminen et al. (2014)

EXPLANATION: This paper identified many areas via real-life surveys wherein visually impaired youth face difficulties. Key facets recognized are independent mobility, social isolation, and barriers to accessing information. The study highlighted that visually impaired students often struggle to navigate through educational environments without external support, maintain social connections, and access, understand and interpret written materials with ease.

TAKEAWAY: VisionPulse aims to address these challenges by incorporating real-time obstacle detection, facial recognition, and optical character recognition (OCR) to assist with navigation, social engagement, and accessing text-based information. These solutions directly mitigate the challenges discussed in this study.

REFERENCE: Anna-Liisa Salminen and Meerit E Karhula. 2014. Young persons with visual

Research Paper 2

TITLE: Mobility Training and Social Participation for Visually Impaired Youth by Ahponen (2008)

EXPLANATION: Ahponen's study surveyed 200 visually impaired youth, with 75% reporting reliance on others for navigation and 70% indicating that this dependency reduced their social participation. Moreover, 65% highlighted the lack of assistive tools as a barrier to independence.

TAKEAWAY: VisionPulse’s real-time obstacle detection and environment recognition systems helps visually impaired students navigate the IIIT Delhi campus independently, hence mitigating the pain point highlighted in the study, which could increase their social participation and reduce dependency.

REFERENCE: H. Ahponen. 2008. Transition to adulthood of severely disabled adolescents: A diverse life course. University of Jyväskylä, Jyväskylä, Finland.

Research Paper 3

TITLE: Computer Use and Independence Among Visually Impaired Adolescents by Pfeiffer et al.

EXPLANATION: Pfeiffer et al. conducted a study involving 180 visually impaired adolescents and 200 sighted counterparts. The study revealed that 85% of visually impaired adolescents used computers daily, compared to just 60% of their sighted peers. This higher usage was attributed to a 78% reliance on technology for communication and 70% for educational purposes.

TAKEAWAY: VisionPulse embraces multimodal AI in offering tools to enhance students' independence in both academic and social settings by helping them gain access to educational materials, communicate effectively, and reduce the effects of this issue.

REFERENCE: P Pfeiffer and M Pinquart. 2013. Computer use of adolescents with and without visual impairment. Technology and Disability 25 (2013), 99–106.

Theme 2:

Direct relevance to VisionPulse’s core engineering functionalities

Research Paper 4

TITLE: VQA via Cross-Modal Retrieval-Augmented Generation

EXPLANATION: The authors formulated a VQA approach that involves employing cross-modal retrieval-augmented generation for relevant model output. They propose a four-step approach named Retrieve (using images as queries to retrieve external knowledge from a database), Augment (incorporating the retrieved knowledge into the question-answering process), Generate (generating answer candidates using a modality-aligned LLM), and Select (choosing the best candidate answer based on specific criteria, such as brevity and accuracy). This four-module pipeline shows great potential in scenarios that require continuously updated and time-sensitive VQA systems

TAKEAWAY: A similar pipeline can be adapted by VisionPulse to assist visually impaired users in navigating specific locations in IIIT Delhi. This framework enables VisionPulse to offer personalized, real-time assistance in location-based tasks for the visually impaired

REFERENCE: P Pfeiffer and M Pinquart. 2013. Computer use of adolescents with and without visual impairment. Technology and Disability 25 (2013), 99–106.

Research Paper 5

TITLE: Tacotron 2: Speech Synthesis in Par with Human Speech

EXPLANATION: Shen et al. propose Tacotron 2, a Google text-to-speech system that generates speech from high-quality text that resembles human speech. The Tacotron 2 integrates a sequence-to-sequence model with WaveNet, reaching a Mean Opinion Score (MOS) of 4.53, almost equivalent to the quality of human speech.

TAKEAWAY: VisionPulse can directly use Tacotron 2's TTS capabilities to give visually impaired students
audio feedback from text-based instructions or environmental data with near-human accuracy, enhancing their
learning and navigation experience

REFERENCE: Jonathan Shen, Ruoming Pang, Ron J. Weiss, et al . 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In 2018
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 4779–4783

Research Paper 6

TITLE: Scaling Efficient Multimodal Large Language Models for Speech Recognition

EXPLANATION: Li et al. of Google proposed a scalable multimodal LLM called PaLM-SayCan for real-time speech recognition and multimodal tasks. It achieves a Word Error Rate (WER) of 2.7%, outperforming previous speech recognition systems.

TAKEAWAY: PaLM-SayCan offers, within the context of VisionPulse, robust speech-to-text (STT) capabilities,
enabling the system to capture user commands and interpret spoken input into actionable responses with high
accuracy, ensuring reliable operation.

REFERENCE: et al. Li, Y. 2022. Scaling Autoregressive Models for Real-time Speech and Multimodal Tasks. arXiv preprint arXiv:2204.02311 (2022).

Research Paper 7

TITLE: Real-time Object Detection Using YOLOv5 by Glenn et al.

EXPLANATION: Glenn Jocher's YOLOv5 model is one of the most widely adopted real-time object detection frameworks, known for its speed and accuracy. YOLOv5 runs at 45 FPS on a standard GPU and has a mean Average Precision (mAP) of 50.5% on the COCO dataset

TAKEAWAY: VisionPulse requires real-time obstacle detection for mobility assistance. This model can be used to
ensure the user receives immediate warnings about nearby obstacles with minimal latency.

REFERENCE: Glenn Jocher et al . 2020. YOLOv5: An open-source implementation of You Only Look Once, version 5. Ultralytics (2020).

Theme 3:

Research papers with useful content in terms of datasets, baselines and evaluation metrics

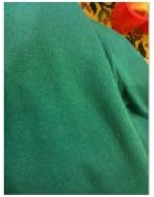
Research Paper 8: Dataset

TITLE: VizWiz Grand Challenge: Answering Visual Questions from Blind People

EXPLANATION: The introduces the VizWiz dataset which consists of over 31,000 visual questions submitted by blind individuals who captured photos using mobile phones and recorded questions about those images. This dataset is unique due to its challenging nature: images are often of poor quality, and some questions are unanswerable because of visual limitations.

Q: Does this foundation have any sunscreen?
A: yes

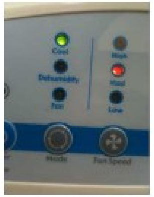
Q: What is this?
A: 10 euros



Q: What color is this?
A: green

Q: Please can you tell me what this item is?
A: butternut squash red pepper soup

Q: Is it sunny outside?
A: yes



Q: Is this air conditioner on fan, dehumidifier, or air conditioning?
A: air conditioning

REFERENCES: Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3608–3617.

Research Paper 9: Metric

TITLE: A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input

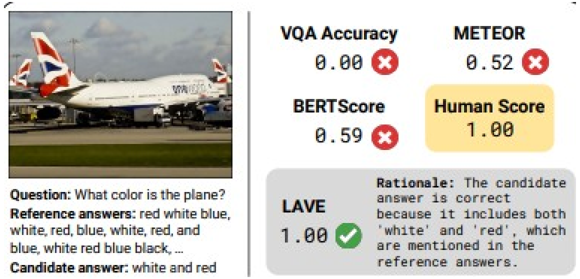
EXPLANATION: The study provided a new metric for evaluation VQA task named Wu-Palmer Similarity Score (WUPS Score). WUPS calculates the similarity between two words based on their longest common subsequence in the taxonomy tree. The WUPS score allows flexible evaluation of object recognition and scene understanding. In real-world environments, objects with semantically similar categories (e.g., "bench" vs. "seat") can receive partial credit, improving usability.

REFERENCES: Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. Advances in neural information processing systems 27 (2014).

Research Paper 10: Metric

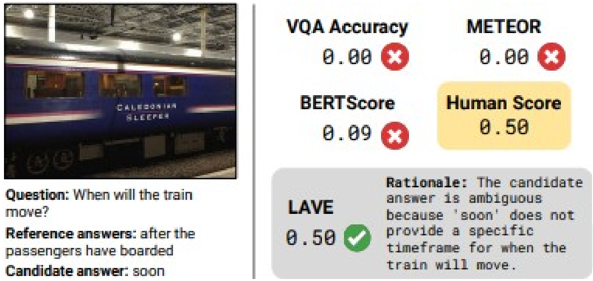
TITLE: Improving Automatic VQA Evaluation Using Large Language Models

EXPLANATION: The study provides a novel technique for evaluating the VQA task named LLM-Assisted VQA Evaluation (LAVE). It uses LLMs like Flan-T5 and GPT to score candidate answers based on their semantic similarity to reference answers instead of using legacy techniques like VQA and soft VQA, outperforming traditional metrics like VQA Accuracy and BERTScore. The study shows that LAVE has higher correlation to human-based evaluation than BERTScore and VQA score.



Question: What color is the plane?
Reference answers: red white blue, white, red, blue, white, red, and blue, white red blue black, ...
Candidate answer: white and red

Rationale: The candidate answer is correct because it includes both 'white' and 'red', which are mentioned in the reference answers.



Question: When will the train move?
Reference answers: after the passengers have boarded
Candidate answer: soon

Rationale: The candidate answer is ambiguous because 'soon' does not provide a specific timeframe for when the train will move.

REFERENCES: Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. Advances in neural information processing systems 27 (2014).

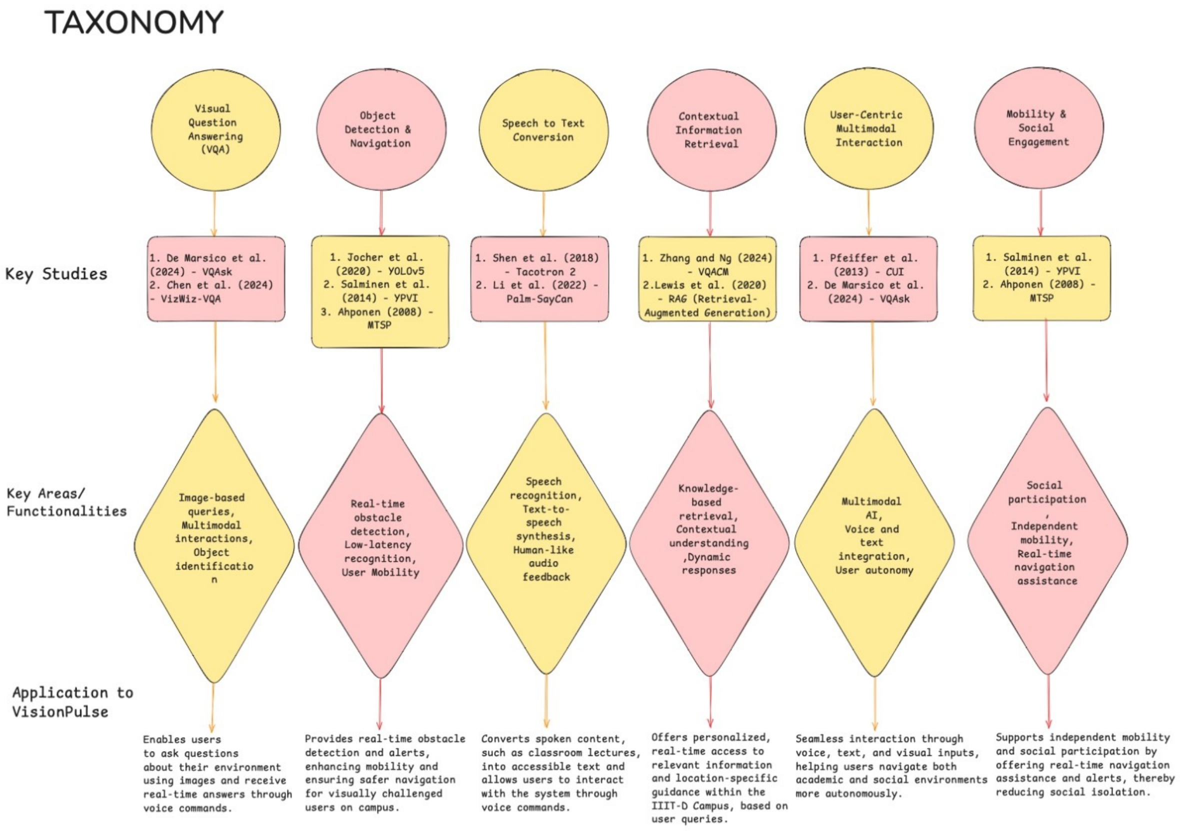
Data Charting - Explanation

Data Charting is diving deeper into the **details of the previous studies relevant to VisionPulse**, by incorporating descriptive-analytical methods to organise into concrete structures and analyse them. The **primary objective** is to **define a mapping** that will aim to shed light **on the academic achievements in the AI-enhanced assistive technologies for the visually challenged people** by **focusing on** essential domains related to our work, namely **object detection, facial recognition, multimodal interaction, speech-to-text and RAG (retrieval augmented generation)**.

It includes the following subsections:->

1. **Descriptive-Analytical Method**
2. **Publication Year and Venue**
3. **Core Contributions and Relevance to VisionPulse**
4. **Taxonomy**
5. **Theoretical Saturation as a Stop Sign**
6. **Key Insights from Data Charting**

Data Charting - Taxonomy



TAXONOMY

	Visual Question Answering (VQA)	Object Detection & Navigation	Speech to Text Conversion	Contextual Information Retrieval	User-Centric Multimodal Interaction	Mobility & Social Engagement
Key Studies	1. De Marsico et al. (2024) - VQAsk 2. Chen et al. (2024) - VizWiz-VQA	1. Jocher et al. (2020) - YOLOv5 2. Salminen et al. (2014) - YPVI 3. Ahponen (2008) - MTSP	1. Shen et al. (2018) - Tacotron 2 2. Li et al. (2022) - Palm-SayCan	1. Zhang and Ng (2024) - VQACM 2.Lewis et al. (2020) - RAG (Retrieval-Augmented Generation)	1. Pfeiffer et al. (2013) - CUI 2. De Marsico et al. (2024) - VQAsk	1. Salminen et al. (2014) - YPVI 2. Ahponen (2008) - MTSP
Key Areas/ Functionalities	Image-based queries, Multimodal interactions, Object identification	Real-time obstacle detection, Low-latency recognition, User Mobility	Speech recognition, Text-to-speech synthesis, Human-like audio feedback	Knowledge-based retrieval, Contextual understanding ,Dynamic responses	Multimodal AI, Voice and text integration, User autonomy	Social participation , Independent mobility, Real-time navigation assistance
Application to VisionPulse	Enables users to ask questions about their environment using images and receive real-time answers through voice commands.	Provides real-time obstacle detection and alerts, enhancing mobility and ensuring safer navigation for visually challenged users on campus.	Converts spoken content, such as classroom lectures, into accessible text and allows users to interact with the system through voice commands.	Offers personalized, real-time access to relevant information and location-specific guidance within the IIIT-D Campus, based on user queries.	Seamless interaction through voice, text, and visual inputs, helping users navigate both academic and social environments more autonomously.	Supports independent mobility and social participation by offering real-time navigation assistance and alerts, thereby reducing social isolation.

A **Taxonomy** (taxonomical view) is used to group the research into different functional aspects of the **VisionPulse** project. The primary aim of the taxonomy is to identify which technologies address specific objectives and functionalities of the **VisionPulse** system.

Conclusion – Technology

Survey helped us oncover the themes under the overarching domains of NLP, CV, Speech Processing
And uncovered specific techniques to tackle the problems.

- **Core Technologies:**
  - Computer Vision (CV), Natural Language Processing (NLP), Speech Processing are foundational to VisionPulse.
  - Object detection with YOLO,
  - VQA multimodal understanding with BLIP and PaLI,
  - Speech recognition with Whisper
  - Retrieval-Augmented Generation (RAG) shows promise for contextually aware responses and using previous and common (institution) or personal context.

- **Current Assistive Technologies:**
  - Products like Seeing AI, NavCog, and Envision Glasses highlighted the potential of using similar tools to improve accessibility.

Conclusion – Approach

Problem	Technology
Real-time object detection and navigation for visually impaired students	YOLO v5
Multimodal interaction and visual question answering (VQA)	BLIP, PaLI
Information organization and summarization	LLMs (llama 3.1 B Instruct Turbo)
Context-aware information retrieval	RAG
Speech-to-text transcription and text-to-speech for acad. engagement	Whisper (STT), Tacotron (TTS)
Location-based guidance for campus navigation	BLIP, PaLI, Bluetooth-based Localization
Email management through voice commands	SMTP Protocols

Shut Down.

Group 1a5b

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI