VisionPulse - Literature Survey Paper

MANAV MITTAL - 2021538, AKASH KUSHWAHA - 2021514, UTKARSH VENAIK - 2021570, LAKSHAY KUMAR - 2021061, SHREYAS KABRA - 2021563, and TANMAY SINGH - 2021569

A clear and well-documented LaTeX document is presented as an article formatted for publication by ACM in a conference proceedings or journal publication. Based on the "acmart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

ACM Reference Format:
Manav Mittal - 2021538, Akash Kushwaha - 2021514, Utkarsh Venaik - 2021570, Lakshay Kumar - 2021061, Shreyas Kabra - 2021563, and Tanmay Singh - 2021569. 2018. VisionPulse - Literature Survey Paper. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Research Questions

In this section, we explore the research questions critical to the development of VisionPulse, an AI-enhanced learning and interaction assistant designed for visually impaired students. The questions focus on the availability of technologies, their applications within educational institutions, the societal dynamics that affect their adoption, and the challenges faced in implementing these solutions.

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

With the recent advancements in artificial intelligence (AI), natural language processing (NLP) and computer vision (CV) have lead to remarkable advancements in assistive technologies for visually impaired individuals. All three techniques together can provide real-time assistance, improve the accessibility of educational resources and enhance learning experience for visually impaired individuals.

- **Vision-based Assistive Systems**: Vision-based techniques which utilize AI and computer-vision techniques, are designed in a manner to interpret the environment and convert the visual stimulus into tactile or auditory feedback, which helps the visually impaired individuals to efficiently interact with their environment in real time. One of the prominent example of this type of system is **Seeing AI** [12] by Microsoft. It utilizes deep learning techniques to detect objects in the susurroundings and respond with auditory descriptions, enabling individuals to receive visual information audibly. Another example of such system is **NavCog** [1], which assists users in indoor navigation with the help the Bluetooth beacons, helping visually impaired individuals to navigate safely.
- **Natural Language Processing for Accessibility**: NLP has also revolutionized how visually impaired individuals access and interact with the information. There are basic tools like **VoiceOver** and **TalkBack** on iOS and Android respectively are now translating on-screen content to spoken words. There are also more advanced

Authors' Contact Information: Manav Mittal - 2021538; Akash Kushwaha - 2021514; Utkarsh Venaik - 2021570; Lakshay Kumar - 2021061; Shreyas Kabra - 2021563; Tanmay Singh - 2021569.

Manuscript submitted to ACM

models like **GPT** and **BERT** which have further improving the accessibility by proving more context aware responses and processing complicated tasks like reading emails and answering questions aloud. These tools can significantly help in educational settings by providing visually impaired students access to large amount of available text based materials.

- **Wearables Technologies**: In recent years wearables have advanced very much due various technological advancements in semiconductor industry. And these have also become valuable tools and have more potential in helping visually impaired individuals. There are devices like **OrCam MyEye** [9] and **Envision Glasses** which utilizes computer vision techniques to recognize faces, reading printed text and providing real-time auditory feedback. These tools offers seamless access to visual information and have potential to empower visually impaired individuals to participate more engagingly in activities.

These technologies and techniques either individually or in combination have transformed the landscape of world significantly for visually impaired individuals by providing them with more autonomy, accessibility and a enhanced experience. These ongoing advancements are pushing the boundaries of what is possible and are helping in overcoming challenges faced by the visually impaired individuals.

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

The application of these tools and technologies for visually impaired individuals depends significantly on the specific environments where these tools are being deployed. We are focusing on IIIT Delhi which is an Educaional institution so in further text we will discuss accordingly. Educational institutions consist of varies spaces, each of those poses different challenges. There are two primary settings in a educational intitution classroom and campuses. These are the two key areas where AI powered systems can significantly improve the learning and overall experience for visually impaired students.

- **Classroom Settings**: Within the classroom settings, the challenges faced by visually impaired are accessing visual materials such as lecture slides and textbooks. To address such issues there are technologies like **LectureSight** and advanced NLP models which are capable pf transcribing and summarizing live video lectures into accessible formats. These AI-powered tools are capable of converting real time spoken content and visual material to structures, readable text, or audio which can provide visually impaired students immediate access to classrom information. This can lead better learning for visually impaired individuals by helping them fully participate in lectured, discussions and note taking.
- **Campus Navigation**: There are several challenges faced by visually impaired students in navigating complex or large campuses. AI-enhanced navigation systems can be crucial in helping them move between various areas with improved ease and independence. For example, **NavCog** [1] uses a combination of different computer vision techniques to provide user a real-time auditory guidance. This can help visually impaired students by directing them through indoor and outdoor areas of the campus such as hallways, classrooms and common areas by offering step-by-step auditory instructions. Additionally, **SeeingAI** [12] by Microsoft can too play an important role in campus navigation by using computer vision techniques to analyze the surroundings and convert the visual information in auditory descriptions. Though SeeingAI is traditionally used for interpreting text, identifying objects and recognizing people, it can also be adapted to help visually impaired students in identifying key landmarks and understanding spatial contexts with campuses, such as reading room numbers, building signs or recognizing familiar faces.

1.3 How are assistive technologies addressing challenges faced by visually impaired students?

AI-based assisive technologies are changing how visually impaired students participate in school by tools lke facial recognition, real-time object detection and speech-to-text systems. These technologies aim to make the learning learning environment more accessible and help the students to become more independent and hence help in overcoming the traditional barriers in education.

- **Real-Time Object Detection:** Visually impaired students often face challenges when navigating their surroundings and understanding the visual information fetched. Real-time object detection such as **YOLO** [14] (You Only Look Once), helps to solve this by identifying the obstacles in their surroundings and reading signs. Theses systems provide instant audio or touch feedback helping students to navigate the nearby objects safely.
- **Contextual AI Systems:** AI systems like **GPT-3** and **PaLM** [6] become helpful for visually impaired students when it comes to learning and processing the information, These robust AI tools helps in tasks like text-to-speech, fetching description of visuals and turning everything into accessible format for the students, For example , during a lecture, these systems can be used to generate a summary or transcriptions in the real time so that students can follow along with the lecture independently. This actually reduces reliance on the human support for any assistance in the learning process.

Collectively, both real-time object detection and Robust AI tools like GPT-3 have the capability to improve both accessibility and the learning experience for visually impaired students.

1.4 How do different societal levels engage with assistive AI for visually impaired people?

Engagement with assistive AI for visually impaired people varies across different levels of society and each one plays an important role in integrating the AI technologies into the educational settings.

- **Individual Level:** At the most personal level, visually impaired individuals leverage assistive AI to overcome day-to-day barriers and enhance their autonomy in social, professional, and personal activities. For instance, tools like Be My AI allow individuals to engage meaningfully with social media by providing detailed descriptions of images, enabling them to navigate online spaces that were previously challenging due to the absence of accessible visual content.
- **Community Level:** Communities, including family, friends, and social networks, interact with assistive AI to better communicate and engage with visually impaired individuals. **Be My AI** [3] , for instance, helps sighted individuals reduce the cognitive load of describing images and situations, bridging the communication gap between the visually impaired and sighted communities. By providing accurate descriptions, Be My AI alleviates the pressure on sighted people to offer perfect visual descriptions, thus facilitating more meaningful interactions.
- **Institution Level:** At an institutional level, assistive AI is being integrated into educational, healthcare, and social service systems to provide enhanced support for visually impaired individuals. Institutions such as universities, healthcare facilities, and public organizations engage with AI to create accessible environments that foster inclusion. AI tools are used to facilitate accessible learning materials, guide independent navigation within institutional spaces, and support broader social participation through access to visual information. Moreover, organizations are starting to integrate AI-based accessibility tools into their service offerings.

1.5 How do educational institutions need to adapt their organizational structures to integrate AI for accessibility?

- **Upgraded Infrastructure:** Schools and universities must improve their digital and physical infrastructure to support AI tools. This could include developing AI-powered learning platforms that work well with assistive technologies like screen readers, voice recognition, and AI-based captioning tools. Additionally, smart navigation systems or personalized AI assistants can help visually impaired students move around campus more easily.
- **Staff Training:** Teachers and staff need training on how to use AI tools to support accessibility and how to integrate them into their teaching methods. This includes learning the technical side of AI tools as well as gaining an understanding of how to create an inclusive environment. Staff should know how to adapt their teaching materials and strategies to meet the diverse needs of students.
- **Adapting the Curriculum:** Educational programs should incorporate AI to ensure visually impaired students can fully participate. Learning materials should be designed with accessibility in mind, using AI to create accessible text, describe visual content, and make real-time adjustments based on students' needs. Schools should also explore ways for visually impaired students to use AI tools during exams.

1.6 Gaps in Research

Despite advancements in assistive technologies, there are still notable gaps in the research and development of AI systems for visually impaired students. These gaps highlight areas where further innovation can be done to ensure that assistive technologies are not only effective but also accessible and practical for everyday use in educational settings.

- **Affordability:** One of the biggest challenges is the very high cost of the AI Assistive technologies which limits the accessibility for the visually impaired individuals specifically in the low and middle income countries, on other hand there are some open source solutions available like **OrCam** [9] or **Envision Glasses** but they lack the advanced features. Research needs to focus on developing affordable and more accessible solutions that also provide comparable functionality to these high-end tools to ensure that financial barriers do not prevent anyone from benefiting from AI-Assistive tools.
- **Non-Intrusive Design:** The recent AI powered assistive devices become sometimes intrusive for distracting users, for instance the wearable technologies like smart glasses may be uncomfortable to be weared for the long period and may draw unwanted attention. There is a need for the further research into the designing the assistive technologies which are more user-friendly and comfortable to use for long periods. This includes creating those devices which are minimally invasive and provide maximum support and allows visually impaired students to use them effortlessly in everyday educational tasks.
- **Cultural and Linguistic Adaptability:** Many AI assistive tools and technologies are developed with a focus on the specific language and cultural contexts, which limits their effectiveness and applications for users of different regions of the world, For example voice recognition and text-to-speech systems often struggle with the different accents of speaking a particular language. This highlights a need for the research that addresses the cultural and linguistic diversity of the users from different regions for ensuring that assistive technologies can adapt to various environments and specific requirement across different regions of the world.

2 Identification and Selection of Relevant Studies

This section discusses our project's relevant studies, projects, and research papers. This comprehensive literature review encompasses an in-depth analysis of various documents drawn from scholarly databases like Google Scholar, ACM Digital Library, IEEE Explore, and others. The research focuses on vision-based AI, multi-modal models, speech-to-text systems, visual question-answering, navigation, and object detection, ensuring that VisionPulse is built upon cutting-edge technological advancements.

2.1 Scholarly Databases Used

The papers discussed in this section have been sourced from three primary scholarly databases stated below. Each of them is a comprehensive library with a huge database of high-quality, peer-reviewed studies that cover a broad range of domains, including vision-based AI assistive technologies.

- **Google Scholar**
- **ACM Digital Library**
- **IEEE Xplore**
- **arXiv**

2.2 Keyword Searching Employed

We employ targeted keyword searches in the databases to identify relevant studies and cover a broad spectrum of appropriate technologies. The following keywords were used to capture the range of technologies pertinent to our project:

- *"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"*
- *"obstacle detection using computer vision"*
- *"visual question answering for visually impaired"*
- *"spatial context-awareness systems"*
- *"RAG-based systems using voice technology"*
- *"real-time navigation using LLMs"*
- *"Retrieval Augmented Generation (RAG)"*
- *"speech-to-text conversion"*
- *"visual captioning"*
- *"educational inclusivity through AI"*
- *"permanent context knowledge base"*
- *"rolling context knowledge base"*
- *"parameter-efficient fine-tuning in LLMs"*

2.3 Selected Relevant Studies

The following studies were selected based on their direct relevance to VisionPulse's core functionalities, including visual question answering (VQA), navigation, obstacle detection, speech-to-text transcription, and contextual information retrieval. They are organised based on themes:

2.3.1 ***Multimodal Applications Designed For Inclusivity.*** According to recent surveys, applications capable of working with multiple modalities—voice, text, and haptics—have proven beneficial for people with vision problems. These systems offer users flexibility to interact with them, making them accessible and aligned with VisionPulse's goals.

- **VQAsk - A Multimodal Android GPT-based Application by De Marsico et al. (2024) [5]**
  The authors discuss their developed application, VQAsk, available on Android, designed to assist visually impaired users by answering questions about their environment through speech-based interaction and haptic feedback. The app employs MiniGPT-4, a vision-language model, and uses automatic image segmentation for object identification. The system allows users to interact with their surroundings through voice commands and enables visual question answering through images. The dataset used in this study, VisWiz, consists of 31,000 visual questions paired with ten answers as ground truth, making it robust for training and fine-tuning the model.
  Relevance: VisionPulse can adopt a similar approach by incorporating functionality akin to MiniGPT-4 and leveraging the VisWiz dataset for training and improving our architecture. The dataset can help train VisionPulse on everyday visual tasks, mainly when IIIT Delhi-specific images are unavailable, such as for generic cases like checking expiry dates. Moreover, the survey of visually impaired participants establishes a preference for voice-based interactions, underscoring the importance of such functionality in VisionPulse for the target students at IIIT Delhi.

2.3.2 ***Visual Question Answering for Assistive Technologies.*** Visual Question Answering (VQA) is a critical field for assistive technologies aimed at visually impaired individuals. VQA systems answer natural language questions about images, which can provide context and understanding for those with limited or no vision.

- **Visual Question Answering for Visually Impaired People (VizWiz-VQA) by Chen et al. (2024) [4]**
  This model uses the VizWiz-VQA dataset, which consists of visual questions asked by visually impaired users. This dataset differs from others in that it is very close to real-world scenarios since the images and answers to questions are taken by the visually impaired people themselves, and the questions are answered based on real-life questions they may have. The study highlights the unique challenges that arise from this authentic use case. Since people face challenges in taking the right pictures, the dataset has various issues like blurred images, which may be encountered in real life.
  Relevance: Incorporating insights from real-world datasets like VizWiz can help VisionPulse better serve visually impaired students by improving the system's ability to interpret and respond to complex visual and textual inputs, even in challenging conditions since the dataset itself has been made by the people who would eventually use it. This dataset can be used for benchmarking of our dataset and also model training for edge case scenarios and full-proofing.

2.3.3 ***Improving VQA Evaluation Metrics with LLMs.*** Traditional VQA evaluation relies on exact string matching, which struggles with open-ended questions. Recent advances propose using large language models (LLMs) for more accurate, human-aligned evaluations.

- **Wu-Palmer Similarity Score (WUPS Score) by Malinowski et al.** [10]
  The WUPS score evaluates system-generated answers by using Fuzzy Set theory and Wu-Palmer Similarity to account for semantic fuzziness between classes. It measures the similarity between predicted and actual answers, awarding partial credit when answers are semantically close. This is particularly valuable when exact matches are rare, but approximations are still informative. WUPS penalizes both underestimation and overestimation, ensuring balanced evaluations and providing a more realistic assessment of performance, especially in tasks like visual question answering where ambiguity in data is common.
  Relevance: The WUPS score allows flexible evaluation of object recognition and scene understanding, crucial for assisting visually impaired users. In real-world environments, objects with semantically similar categories (e.g., "bench" vs. "seat") can receive partial credit, improving usability. It also helps handle complex scene descriptions, enabling VisionPulse to adapt to diverse and ambiguous inputs while offering reliable assistance in daily tasks.
- **LLM-Assisted VQA Evaluation (LAVE) by Mañas et al. (2024) [11]**
  Developed LAVE evaluation metric, which 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. It was demonstrated that the LAVE metric has higher correlation to human-based evaluation than BERTScore and VQA score.
  Relevance: LAVE's ability to handle open-ended questions is highly relevant for VisionPulse, improving real-time VQA evaluation for visually impaired students by offering more nuanced and accurate feedback. This shows that there is a need to test the model keeping in mind the ultimate goal of usability and human-friendliness.

2.3.4 ***Understanding the Challenges Faced By Visually Impaired Youth.*** Young people with visual impairments face considerable barriers in different facets of life, particularly in education, mobility, and social interactions. Understanding these challenges is crucial for our application so that we can aim to mitigate them and understand and solve real-life issues.

- **Young Persons with Visual Impairment: Challenges of Participation by Salminen et al. (2014) [15]**
  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 though educational environments without external support, maintain social connections, and access, understand and interpret written materials with ease.
  Relevance: 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.
- **Mobility Training and Social Participation for Visually Impaired Youth by Ahponen (2008) [2]**
  The study explores the dependency of visually impaired youth on others for mobility and social participation due to the lack of assistive tools.
  Relevance: 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.

- **Computer Use and Independence Among Visually Impaired Adolescents by Pfeiffer et al.[13]**
  The study revealed that the computer usage of visually impaired adolescents is higher than that of their counterparts, and the research found it was all because of their dependency on technology for day-to-day communication and education.
  Relevance: 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.

2.3.5 ***State-of-the-Art Multimodal Technologies for Assistive Systems.*** Recent advancements in multimodal technologies and state-of-the-art models have transformed assistive systems for visually impaired individuals, especially in text-to-speech, speech-to-text, and real-time object detection. These systems are crucial for building comprehensive solutions like VisionPulse that require high accuracy and reliability.

- **Tacotron 2: Speech Synthesis in Par with Human Speech [16]**
  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.
  Relevance: 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.

- **Scaling Efficient Multimodal Large Language Models for Speech Recognition [8]**
  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.
  Relevance: 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.

- **Real-time Object Detection Using YOLOv5 [7]**
  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.
  Relevance: 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.

2.3.6 ***Retrieval-Augmented Generation (RAG) for Assistive Systems.*** Standard generative models rely solely on their training dataset; in contrast, RAG combines a generative model with an external knowledge base, which helps generate contextually accurate and more personalized responses for the user.

- **VQA via Cross-Modal Retrieval-Augmented Generation** [17]
  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.
  Relevance: 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.

3 Data Charting

In this section, we dive 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). To conclude this section, we present a **taxonomy** (taxonomical view), to categorise the studies, providing a clearer picture on their contributions to different aspects within our project.

3.1 Descriptive-Analytical Method

In this sub-section, we present to you the selected studies in accordance with our project, by synthesizing them using key parameters such as **Publication year**, **Venue of Publication**, **Core Contributions**, and it's **Relevance** to VisionPulse.

3.2 Publication Year and Venue

The most recent advancements in the field of AI are prioritised by keeping a track of its evolution and assistive technologies on the basis of the Publication Year. For instance, the recent studies from 2024 on VQAsk and LAVE, demonstrates an application for Visual Question Answering using LLMs and haptic feedbacks, and the proposal of an evaluation metric having a high degree of correlation with human-based evaluation. The Venue, International Conference on Advanced Visual Interfaces and ArXiv, provides credibility to the work, focusing on previous studies in the related domain in technical peer-reviewed journals and high-profile conferences.

3.3 Core Contributions and Relevance to VisionPulse

The core contributions and relevance of each study is evaluated based on their impact on key components such as object recognition, multimodal interaction, speech-to-text transcription, and real-time obstacle detection. By emphasizing the practical aspects of each model, we ensure that VisionPulse adopts the most appropriate technologies to achieve its objectives.

- **VQAsk**: The study introduces the **VQAsk** application, which leverages **MiniGPT4-a** and automatic image segmentation to assist visually challenged users in answering questions about their environment. The use of **Visual Question Answering (VQA)** through speech and haptic feedback is directly relevant to VisionPulse's goal of enabling users to interact with their surroundings via voice commands and image-based queries. By integrating these technologies, VisionPulse can enhance its capability to process real-time, image-based inputs for dynamic user interaction.
- **VizWiz-VQA**: The **VizWiz** dataset, which is compiled from real-life scenarios involving visually challenged users, provides a rich resource for VisionPulse's **model training** and **benchmarking**. This dataset captures practical challenges such as blurred images and varying lighting conditions, making it ideal for training VisionPulse's models to handle edge cases. By utilizing this data, VisionPulse ensures robustness and adaptability in real-world environments, leading to a more reliable and practical solution for visually challenged users.
- **WUPS Score**: The **WUPS Score** metric offers a more profound method for evaluating model performance by awarding partial credit for semantically similar answers. This is particularly relevant for VisionPulse's **object recognition** and **scene understanding** components, where exact matches may not always be found, but semantically close answers may still provide useful assistance. Incorporating the WUPS Score ensures that VisionPulse's evaluation methods are flexible and reflective of real-world complexity, thereby, enhancing the system's overall reliability.
- **LAVE**: The **LAVE** evaluation metric, which leverages the use of models like FLAN-T5 and GPT, significantly improves the accuracy measurements in **Visual Question Answering (VQA)** tasks. This metric's ability to correlate closely with human-based evaluations makes it an essential tool for VisionPulse's **real-time VQA evaluation**. By adopting the LAVE metric, VisionPulse can offer more nuanced feedback for open-ended questions, thereby improving the system's responsiveness and user experience for visually challenged individuals.
- **YPVI**: The study identifies critical challenges faced by visually challenged individuals, primarily, the difficulties with **independent mobility**, **social isolation**, and **access to information**. VisionPulse addresses these challenges by incorporating **real-time obstacle detection**, **facial recognition**, and **optical character recognition (OCR)** technologies. These features ensure that users can navigate safely, engage socially, and access important information with ease, aligning with VisionPulse's goal of enhancing independence and quality of life for visually challenged users.
- **MTSP**: The research discovers the mobility and social participation barriers faced by visually challenged users due to a lack of adequate assistive tools. VisionPulse aims to overcome these challenges by integrating **real-time obstacle detection** and **environment recognition** technologies, empowering users to navigate spaces more independently. By addressing the specific mobility needs of visually challenged individuals, VisionPulse contributes to improving their ability to engage in social and physical environments significantly.
- **CUI**: The study examines the dependence of visually challenged students on technology for communication and education, emphasizing the need for tools that enhance the degree of independence. VisionPulse's **multimodal AI** approach is designed to reduce this dependence by offering **voice**, **text**, and **visual feedback**, enabling students to communicate and learn more independently. By supporting academic and social engagement, VisionPulse aligns with the study's goal of promoting self-reliance in educational domain.
- **Tacotron 2**: The **Tacotron 2**, a Google text-to-speech system, generates highly realistic, human-like speech from textual data, making it an ideal solution for providing **audio feedback** in VisionPulse. It can be significantly

useful in providing navigation instructions or responding to environmental data. Moreover, the Tacotron 2's natural-sounding voice output enhances the user experience, especially for visually challenged individuals who rely on audio cues for learning and interaction.

- **SEMLL**: The **Palm-SayCan** model introduced in the study offers advanced **speech-to-text** and **multimodal task** capabilities, thus, making it a robust solution for VisionPulse's real-time **speech recognition** objective. By accurately capturing and interpreting spoken commands, VisionPulse can deliver real-time responses and facilitate seamless interaction between users and their environment. This speech-to-text functionality is crucial for enabling visually challenged users to communicate effectively in various contexts within the system.
- **YOLOv5**: The **YOLOv5** is one of the most efficient real-time **object detection** frameworks and is a vital cog in VisionPulse's **obstacle detection** component. Its high degree of speed and accuracy allows VisionPulse to provide users with immediate warnings about nearby obstacles, ensuring their safety in dynamic environments. This real-time detection capability is particularly important for visually challenged users navigating unfamiliar or crowded spaces, where quick responses are critical for avoiding hazards and collisions.
- **VQACM**: The **RAG-based VQA pipeline** proposed in the study is highly relevant for VisionPulse's real-time **Visual Question Answering (VQA)** functionality. By utilizing the Retrieve, Augment, Generate, and Select (RAG) method, VisionPulse can offer personalized, context-aware assistance to visually challenged users to navigate specific environments within the IIIT-D campus.

3.4 Taxonomy

In this subsection, we introduce a **taxonomy** (taxonomical view) 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.

3.5 Theoretical Saturation as a Stop Sign

In this section, **theoretical saturation** refers to the point where further literature review provides very little novel insights into our work. For instance, after reviewing multiple studies on object detection, such as YOLOv5 (Jocher et al., 2020) and real-time obstacle detection technologies from Salminen et al. (2014) and Ahponen (2008), we can conclude that these technologies are mature enough to be used in VisionPulse. Similarly, after examining core technologies for visual question answering, including VQAsk (De Marsico et al., 2024), VizWiz-VQA (Chen et al., 2024), LAVE (Mañas et al., 2024), and VQACM (Zhang and Ng, 2024), as well as speech transcription systems like Tacotron 2 (Shen et al., 2018) and Palm-SayCan (Li et al., 2022), theoretical saturation will signal when to stop further research unless new and highly significant developments arise.

3.6 Key Insights from Data Charting

These are some of the key findings we discovered from Data Charting:

- **Comprehensive Integration of Assistive Technologies**: The studies cover a wide range of technologies crucial for VisionPulse, including **multimodal interaction** (Pfeiffer et al., 2013), **real-time obstacle detection** (Jocher et al., 2020; Salminen et al., 2014; Ahponen, 2008), and **speech transcription** (Shen et al., 2018; Li et al., 2022). Collectively, these technologies create a comprehensive platform that helps visually challenged users with navigation, communication, and understanding of their environment.

Table 1. Descriptive Analytical Method

Study	Publication Year	Venue of publication	Core Contributions	Relevance to VisionPulse
De Marsico et al. (2024) - VQAsk	2024	Proceedings of the 2024 International Conference on Advanced Visual Interfaces	A speech and haptic feedback based interactive application, VQAsk, deployed on Android to be used by visually challenged users to answer questions about their environment. The app uses MiniGPT-a and automatic image segmentation for object identification to enable question answering through images.	MiniGPT4-a and VisWiz dataset can be used for model training and improving our model architecture. Dataset is ideal for generic model training and the use of voice commands for question answering makes it an ideal choice.
Chen et al. (2024) - VizWiz-VQA	2024	Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.	Uses a dataset modelled on real-life use cases by gathering data from visually challenged users, themselves. Captures real life scenarios such as blurry images and contributes as a high quality dataset in the dataset pool.	The dataset, VizWiz, is ideal for model benchmarking, training for edge case scenarios and full-proofing since it contains real world data captured by the visually challenged users themselves.
Mañas et al. (2024) - LAVE	2024	ArXiv	Developed LAVE evaluation metric, based on FLAN-T5 and GPT to achieve high correlation to human-based evaluations by outperforming traditional metrics such as VQA Score and BERTScore.	Highly relevant to handle open-ended questions, improving real-time VQA evaluation for visually challenged users by offering more nuanced and accurate feedback.
Shen et al. (2018) - Tacotron 2	2018	IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE	Generated human-like speech using Google's text-to-speech system that generates speech from high-quality textual data.	Ideal to give audio feedback to Visually challenged students from text-based instructions or environmental data, enhancing their learning and navigation experience.
Li et al. (2022) - SEMLL	2022	ArXiv	Proposed Palm-SayCan, a scalable multimodal LLM, for real-time speech recognition and multimodal tasks.	Provides a robust, Speech-to-text capable system to capture user commands and interpretation of spoken input into actionable sequences.
Glenn Jocher et al. (2020) -YOLOv5	2020	Ultralytics	Created YOLOv5, the most renowned and widely-adopted real time object detection framework known for it's speed and accuracy.	Ideal for real-time obstacle detection for mobility assistance by throwing immediate warnings about nearby obstacles with minimal latency.
Liyang Zhang and Youyang Ng. (2024) - VQACM	2024	-	Developed a four-module pipeline for VQA using RAG, namely Retrieve, Augment, Generate and Select, which is useful in scenarios requiring continuously updated and time-sensitive VQA systems.	Ideal Pipeline for VisionPulse to offer personalised, real-time assistance for visually challenged students to navigate specific locations of IIIT-D.

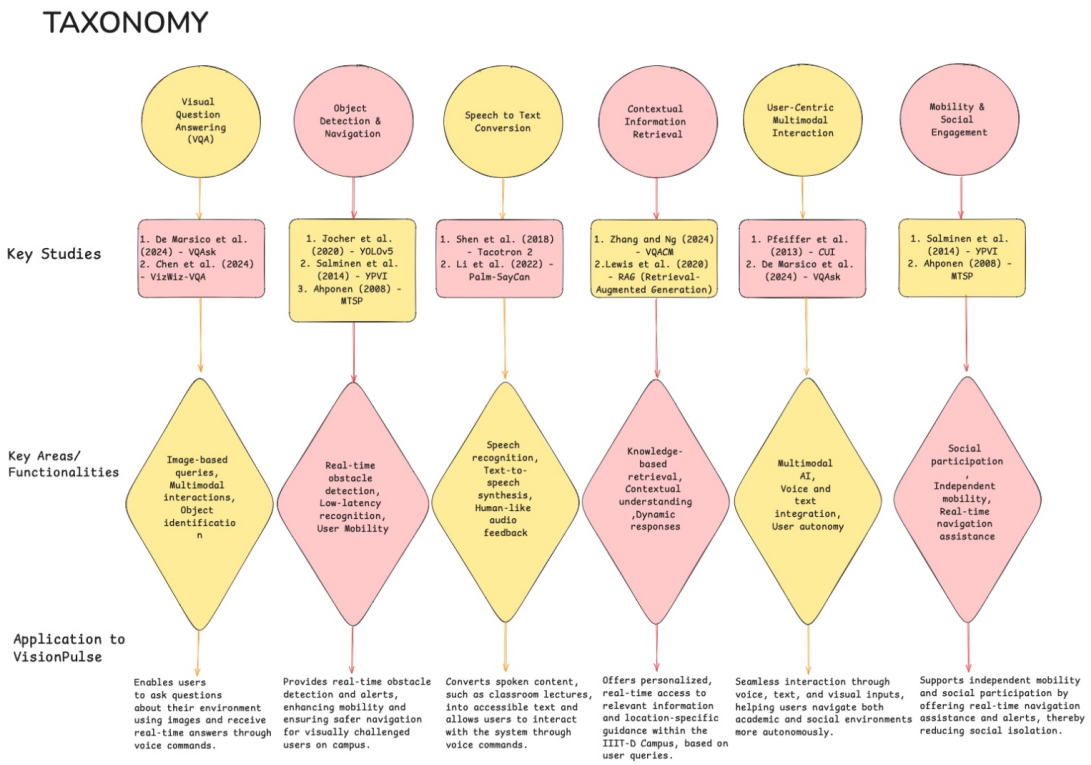

Fig. 1. Taxonomy

- **Advanced Visual Question Answering (VQA)**: Models like **VQAsk** (De Marsico et al., 2024), **VizWiz-VQA** (Chen et al., 2024), **LAVE** (Mañas et al., 2024), and **VQACM** (Zhang and Ng, 2024) offer advanced solutions for **visual question answering**, enabling users to ask questions about their surroundings and receive accurate, real-time answers. These models form the foundational stone for the VisionPulse's **Q&A and navigation systems**, helping users better interact with their environment.
- **Real-Time Object Detection for User Safety**: Object detection technologies such as **YOLOv5** (Jocher et al., 2020) and obstacle detection systems (Salminen et al., 2014; Ahponen, 2008) are key to VisionPulse's **safety features**. These systems provide fast, accurate detection of obstacles, allowing visually challenged users to navigate safely, especially in a complex or unfamiliar environment.
- **Efficient Information Retrieval and Contextual Assistance**: The **RAG-based pipeline** (Zhang and Ng, 2024) enhances VisionPulse's ability to deliver **context-aware information retrieval**, thereby, offering real-time and personalized assistance for tasks such as **campus navigation** and **academic support**.
- **Proficient Speech Recognition and Audio Feedback**: **Tacotron 2** (Shen et al., 2018) and **Palm-SayCan** (Li et al., 2022) provide advanced **speech-to-text** and **text-to-speech** capabilities that are essential for VisionPulse's **multimodal interaction system**. These technologies allow users to issue commands and receive clear, human-like audio responses, ensuring a smooth and seamless experience.

- **Relevance to Real-World Scenarios**: The studies like **VizWiz** (Chen et al., 2024) and **MiniGPT4-a** from **VQAsk** (De Marsico et al., 2024) are based on real-world data and challenges faced by visually challenged individuals. This makes them highly relevant for VisionPulse, which aims to address everyday issues by improving its model's functionality in practical scenarios.
- **Focusing on User Demands**: The studies address the broader challenges faced by visually challenged individuals, including **mobility difficulties** and **social isolation** (Salminen et al., 2014; Ahponen, 2008), as well as **reliance on technology** in educational settings (Pfeiffer et al., 2013). VisionPulse leverages the use of technological solutions that promote greater independence and autonomy, both in social and academic environments.

4 Conclusion

This structured analysis of literature helped us to identify technologies, methodologies, and gaps related to assistive systems, particularly aimed at aiding visually impaired students. Below is a summarized view of the insights gathered and their alignment with the development of the VisionPulse project.

4.1 Familiarization

The scope of the project delved into essential technologies such as:

- **Computer Vision (CV)**, **Natural Language Processing (NLP)**, and **Speech Processing**: These fields were identified as critical for processing visual and audio inputs to aid visually impaired individuals in navigation, object detection, and information retrieval.
- Models and frameworks including:
  - **YOLO** for object detection,
  - **BLIP** and **PaLI** for multimodal understanding,
  - **Whisper** for speech-to-text systems.
- Technologies such as **Retrieval-Augmented Generation (RAG)** and real-time transcription models showed significant potential for context-aware information processing, which are foundational for the project's assistive capabilities.

Several assistive products for the visually impaired are available in the market. **Seeing AI** provides real-time object detection and scene descriptions via auditory feedback. **NavCog** assists with indoor navigation using Bluetooth beacons, while wearable devices like **OrCam MyEye** and **Envision Glasses** use computer vision to read text, recognize faces, and offer auditory descriptions. These technologies enhance independence and accessibility, enabling users to navigate and interact with their environments more effectively.

4.2 Coding

VisionPulse is structured around a modular API-based architecture to ensure seamless communication between different components. The core technologies include:

- **YOLOv5** for real-time object detection, providing visually impaired students with immediate feedback for obstacle avoidance in a dynamic campus environment.
- **BLIP** and **PaLI** models for visual question answering (VQA) and multimodal interactions, which will be fine-tuned with IIIT Delhi-specific data to offer personalized responses about the user's surroundings.

- For **speech-to-text** and **text-to-speech**, VisionPulse incorporates **Whisper** for lecture transcription and voice command recognition, and **Tacotron 2** for delivering natural-sounding audio feedback.
- **Retrieval-Augmented Generation (RAG)** pipelines provide contextual information retrieval, allowing users to query previously stored data and receive contextually accurate responses based on both real-time inputs and stored context.

The system integrates computer vision, natural language processing, and speech synthesis modules using custom APIs. This modular architecture ensures fluid, real-time interaction between the user and the environment, enhancing the system's ability to assist visually impaired students.

4.3 Use Cases and Features

The VisionPulse project offers a range of features that address the specific needs of visually impaired students within a university campus:

- **Real-time Object Detection and Navigation**: VisionPulse employs YOLOv5 to detect obstacles and provide real-time feedback, ensuring safe navigation around the campus. This feature is critical in complex environments, preventing accidents by alerting users to hazards in their immediate surroundings.
- **Multimodal Interaction and Q&A**: Users can issue voice commands, receive visual feedback, and access answers through text or audio. BLIP, PaLI, and RAG are used for visual question answering and contextual query handling, providing detailed responses to queries such as "What's in front of me?" or "Where am I?"
- **Speech-to-Text and Text-to-Speech Capabilities**: VisionPulse supports real-time transcription of lectures via Whisper, helping students follow along and engage in classroom activities. The Tacotron 2 model delivers high-quality audio responses, providing users with accessible access to text-based academic materials.
- **Context-Aware Navigation**: The system integrates Bluetooth-based localization and RAG pipelines to provide location-based guidance. Users can receive real-time directions, allowing them to navigate the campus autonomously.
- **Mailbox Integration**: VisionPulse integrates with the student's mailbox using SMTP protocols, enabling users to read and write emails through voice commands. This functionality promotes better academic and social interaction within the university community.

4.4 Final Thoughts

The findings from this literature survey provide a comprehensive blueprint for the development of VisionPulse. With potential technologies to use and baselines, VisionPulse has the potential to revolutionize how visually impaired students interact with their environment, making educational spaces more accessible and navigable.

References

[1] Dragan Ahmetovic, Cole Gleason, Chengxiong Ruan, Kris Kitani, Hironobu Takagi, and Chieko Asakawa. 2016. NavCog: a navigational cognitive assistant for the blind. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Florence, Italy) *(MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 90–99. https://doi.org/10.1145/2935334.2935361

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

[3] Oliver Bendel. 2024. How Can Generative AI Enhance the Well-being of Blind? arXiv:2402.07919 [cs.HC] https://arxiv.org/abs/2402.07919

[4] Chongyan Chen, Samreen Anjum, and Danna Gurari. 2022. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19098–19107.

[5] Maria De Marsico, Chiara Giacanelli, Clizia Giorgia Manganaro, Alessio Palma, and Davide Santoro. 2024. VQAsk: a multimodal Android GPT-based application to help blind users visualize pictures. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*. 1–5.

[6] Mohammad Fahes, Christophe Kervazo, Jérôme Bobin, and Florence Tupin. 2022. Unrolling PALM for sparse semi-blind source separation. arXiv:2112.05694 [astro-ph.IM] https://arxiv.org/abs/2112.05694

[7] Glenn Jocher et al. 2020. YOLOv5: An open-source implementation of You Only Look Once, version 5. *Ultralytics* (2020). https://github.com/ultralytics/yolov5

[8] et al. Li, Y. 2022. Scaling Autoregressive Models for Real-time Speech and Multimodal Tasks. *arXiv preprint arXiv:2204.02311* (2022). https://arxiv.org/abs/2204.02311

[9] Yimin Lin, Kai Wang, Wanxin Yi, and Shiguo Lian. 2019. Deep Learning based Wearable Assistive System for Visually Impaired People. arXiv:1908.03364 [cs.RO] https://arxiv.org/abs/1908.03364

[10] 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).

[11] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving Automatic VQA Evaluation Using Large Language Models. arXiv:2310.02567 [cs.CV] https://arxiv.org/abs/2310.02567

[12] Microsoft. 2017. Seeing AI - Talking Camera for the Blind. *NA* (2017).

[13] JP Pfeiffer and M Pinquart. 2013. Computer use of adolescents with and without visual impairment. *Technology and Disability* 25 (2013), 99–106.

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640 [cs.CV] https://arxiv.org/abs/1506.02640

[15] Anna-Liisa Salminen and Maarit E Karhula. 2014. Young persons with visual impairment: Challenges of participation. *Scandinavian Journal of Occupational Therapy* 2014 (2014), 1–10. https://doi.org/10.3109/11038128.2014.899622

[16] 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. https://doi.org/10.1109/ICASSP.2018.8461368

[17] Liyang Zhang and Youyang Ng. 2024. Visual Question Answering via Cross-Modal Retrieval-Augmented Generation of Large Language Model. In *38 (2024).* , 2O1GS301–2O1GS301.