UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS

Computer Vision Driven Assistive Solution for People with Visual Impairment or Blindness

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Mohammad Moeen Valipoor
MSc in Software Engineering

Madrid, 2024

UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS

Doctoral Degree in Software, Systems and Computing

Computer Vision Driven Assistive Solution for People with Visual Impairment or Blindness

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Mohammad Moeen Valipoor
MSc in Software Engineering

Under the supervision of:
Dra. Angélica de Antonio Jiménez
Dr. Julián Cabrera Quesada

Madrid, 2024

Title: Computer Vision Driven Assistive Solution for People with Visual Impairment or Blindness

Author: Mohammad Moeen Valipoor

Doctoral Programme: Software, Systems and Computing

Thesis Supervision:

Dra. Angélica de Antonio Jiménez, Associate Professor, Universidad Politécnica de Madrid (Supervisor)

Dr. Julián Cabrera Quesada, Associate Professor, Universidad Politécnica de Madrid (Supervisor)

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

Acknowledgement

Firstly, I'd like to thank my supervisors, Angélica de Antonio Jiménez and Julián Cabrera Quesada, for their invaluable guidance, constructive feedback, and encouragement throughout this journey. Their expertise and dedication have been critical in shaping this work and advancing my career as a researcher.

Secondly, I am deeply grateful to Gofore Spain S.L. for giving me the opportunity to pursue this industrial doctorate. Their trust, collaboration, and support were critical to this project's success. The experience has been both professionally enriching and personally rewarding, and I appreciate the opportunity to contribute to their innovative vision.

Finally, I'd like to express my heartfelt gratitude to my friends and family in Iran and Spain, who have been an unwavering source of support.

Abstract

This thesis aims to support researchers and developers in creating cost-effective, computer vision-based assistive solutions for people with visual impairment and blindness (P-VI/blindness) in indoor environments.

The primary objective of this research is to propose a framework for designing cost-effective solutions that assist P-VI/blindness in understanding their surroundings and locating objects in indoor environments that addresses the main existing challenges. The secondary objective of this thesis is to apply the proposed framework for the development of a specific cost-effective solution that helps P-VI/blindness with scene understanding and locating objects in their surroundings.

At the beginning, a comprehensive systematic mapping study (SMS) was conducted to understand the advances of computer vision in the field of assistive solutions for scene understanding during the past years. Then a semi-structured interview with eight participants having different levels of visual impairment was performed to better understand user needs.

The SMS combined with the user research served to identify some key challenges in the development of such solutions:

- The integration of user needs and requirements during the design and development process.
- The selection of appropriate technologies (both hardware and software) among the various available options, that align with an intended solution.
- Effectively communicating feedback from the system to users.

Afterwards, a framework was developed that addresses the mentioned challenges and guides researchers and developers in integrating user needs, selecting appropriate technologies, and effectively communicating system feedback to users. The framework consists of:

- The various use cases of an assistive solution for scene understanding in indoor environments.
- The list of functional and non-functional requirements to be fulfilled by an assistive solution for scene understanding in indoor environments.
- A general reference architecture for assistive solutions for scene understanding in indoor environments.

- A guideline for selecting appropriate technologies for the design and development of such solutions.

Moreover, according to the proposed framework, two solutions are presented to fulfil the secondary objectives of the thesis. This thesis adopts two academic and industrial approaches. The academic approach focuses on analyzing and evaluating state-of-the-art technologies, while the industrial approach aims to develop a minimum viable product as a commercial solution.

Two distinct solutions were developed as secondary objectives:

1. AssistDiv: A wearable laptop-based solution utilizing an RGB-D camera for scene understanding and object location assistance. This prototype was used to assess various state-of-the-art technologies and was tested with blindfolded participants.
2. V-ASSISTANT: A smartphone-based minimum viable product developed to explore industrialization potential. This solution underwent preliminary testing with P-VI/blindness users.

In conclusion, this thesis provides a valuable framework and insights for researchers and developers working on assistive technologies for P-VI/blindness. It addresses critical challenges in indoor scene understanding and object location, paving the way for more effective, user-centered solutions that can significantly improve the daily lives of P-VI/blindness.

Resumen

Esta tesis tiene como objetivo apoyar a investigadores y desarrolladores en la creación de soluciones de asistencia rentables y efectivas, basadas en visión por computadora, para personas con discapacidad visual y ceguera (P-DV/ceguera) en entornos interiores.

El objetivo principal de esta investigación es proponer un marco para diseñar soluciones rentables y eficaces que ayuden a las P-DV/ceguera a comprender su entorno y localizar objetos en entornos interiores, abordando los principales desafíos existentes. El objetivo secundario de esta tesis es aplicar el marco propuesto para el desarrollo de una solución específica que ayude a las P-VI/ceguera en la comprensión y la localización de objetos en su entorno.

Al inicio, se realizó un estudio de mapeo sistemático (SMS) exhaustivo para comprender los avances de la visión por computadora en el campo de las soluciones de asistencia para la comprensión del entorno en los últimos años. Posteriormente, se realizaron entrevistas semiestructuradas con ocho participantes que tenían diferentes niveles de discapacidad visual para comprender mejor las necesidades de los usuarios.

El SMS, combinado con las entrevistas semiestructuradas, identificó los principales desafíos en el desarrollo de dichas soluciones:

- La integración de las necesidades y requisitos del usuario durante el proceso de diseño y desarrollo.
- La selección de tecnologías apropiadas (tanto de hardware como de software) entre varias opciones disponibles, que se alineen con la solución prevista.
- Comunicar de manera efectiva la retroalimentación del sistema a los usuarios.

Posteriormente, se desarrolló un marco que cubre los desafíos mencionados y guía a los investigadores y desarrolladores en la integración de las necesidades del usuario, la selección de tecnologías adecuadas y la comunicación efectiva de la retroalimentación del sistema a los usuarios. El marco se compone de:

- Los diversos casos de uso de una solución de asistencia para la comprensión de escenas en entornos interiores.
- La lista de requisitos funcionales y no funcionales que deben cumplir las soluciones de asistencia para la comprensión de escenas en entornos interiores.
- Una arquitectura de referencia general para soluciones de asistencia para la comprensión de escenas en entornos interiores.
- Una guía para la selección de tecnologías apropiadas para el diseño y desarrollo de tales soluciones.

Además, de acuerdo con el marco propuesto, se presentan dos soluciones para cumplir con los objetivos secundarios de la tesis. Esta tesis adopta dos enfoques: uno académico y otro industrial. El enfoque académico se centra en analizar y evaluar las tecnologías más avanzadas, mientras que el enfoque industrial tiene como objetivo desarrollar un producto mínimo viable como solución comercial.

Se desarrollaron dos soluciones distintas como objetivos secundarios:

1. AssistDiv: Una solución portátil basada en una laptop que utiliza una cámara RGB-D para la comprensión de escenas y la asistencia en la localización de objetos. Este prototipo se utilizó para evaluar diversas tecnologías avanzadas y fue probado con participantes con los ojos vendados.
2. V-ASSISTANT: Un producto mínimo viable basado en un smartphone desarrollado para explorar el potencial de industrialización. Esta solución fue sometida a pruebas preliminares con usuarios P-VI/ceguera.

En conclusión, esta tesis proporciona un marco valioso y conocimientos útiles para investigadores y desarrolladores que trabajan en tecnologías de asistencia para P-VI/ceguera. Aborda desafíos críticos en la comprensión de escenas en interiores y la localización de objetos, allanando el camino para soluciones más efectivas y centradas en el usuario que puedan mejorar significativamente la vida diaria de las P-VI/ceguera.

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

CNN	Convolutional Neural Network
DSRM	Design Science Research Methodology
GPS	Global Positioning System
MTCNN	Multi-task Cascaded Neural Networking
P-VI	People with Visual Impairment
SLAM	Simultaneous Localization and Mapping
SMS	Systematic Mapping Study
TOF	Time of Flight

1 Introduction

1.1 Contextualization and Background

The motivation for this thesis stems from the need to develop innovative solutions for assisting the people with visual impairment/blindness (P-VI/blindness) in understanding their surroundings in indoor environments. This work draws inspiration from a previous doctoral thesis directed by Dr. De Antonio—"Computational model for the generation of directions for object location in virtual environments: spatial and perceptual aspects," defended by Graciela Lara in 2016. Her thesis was focused on the problem of generating indications, as messages in natural language, to assist the user in the location of objects. The proposed solution was applicable only in fully virtual environments and was oriented to provide assistance to a sighted user.

Conversely, in this thesis, real spaces, and real objects, will be considered. In terms of developing computerized representations of the environment, considering real locations and things poses an immensely complicated problem. While in virtual environments, a semantic model of the environment can be pre-constructed, and its dynamic update is relatively simple from the interactions and behaviours, in real environments, we rarely have previous computerized models of the environment, and keeping these models updated is difficult because changes in the environment can occur without the system being aware of them.

Additionally, the target end user for this thesis are P-VI/blindness which makes the whole process of system's feedback different in comparison with a system designed for the sighted users. Graciela Lara's thesis focused on the automatic generation of natural language instructions to assist in locating objects. These instructions could take egocentric references (relative to the user's body) or exocentric references (relative to other objects) and relied heavily on visual characteristics (size, color, shape) and spatial relationships between objects. The first step for the generation of instructions involved selecting an appropriate reference object, which was achieved through a computational model of object's perceptual salience combined with other factors such as proximity or user's previous knowledge. Next, spatial relationships between the reference object and the target object were determined, resulting in instructions such as "the bracelet is to the left of the yellow vase that is on the round table".

However, these types of instructions would not be useful for users with visual impairments, so new criteria and algorithms for generating instructions would be needed.

According to a World Health Organization (WHO) study, there are approximately 285 million persons with visual impairment (P-VI) worldwide, with 39 million completely blind [1]. Many of these people struggle with some of their daily tasks. These tasks include navigating an environment, identifying obstacles in their path, and identifying objects in their surroundings.

Furthermore, forming a mental map is difficult for them due to their visual limitations. Cognitive or mental maps are mental models used to learn, understand, simplify, and explain human interactions with their surroundings, which include object locations, observations, routes, spatial relations, and so on [2]. Various solutions [3], [4], [5] have been created and developed over the years to address the challenges that P-VI/blindness face. For example, providing information about their surroundings in the form of audio or tactile feedback can aid in the formation of mental maps [6]. Moreover, by analyzing and describing image content from the users' surroundings, scene descriptions can aid in developing a cognitive understanding of surroundings, thus easing navigation for those with visual impairments.

The process of describing the image content is a complicated task, since it combines techniques from computer vision and natural language processing in order to analyze and understand images and generate descriptions. Besides, for having a more vivid description, it is necessary to depict the objects' positions in the three-dimensional space and express the relationships amongst these objects, which adds an extra layer of complexity to this task. It is also difficult to determine which parts of the information about the scene are most relevant to the user and should be included in the description, leaving out others that would make the description too long. There has been previous work on image description [7], [8] mainly for sighted users but the process becomes even more challenging when describing a scene or an image to a person who is blind.

Many researchers have been working on assistive solutions that use technologies such as RGB-D cameras, ultrasonic sensors, optical beacons, LiDAR, RFID tags, or WiFi access points to make everyday tasks easier for the P-VI/blindness. These new solutions have many advantages over traditional ones, such as white canes and guide dogs. White canes, for instance, cannot be used to detect obstacles that are farther away, and they are only useful for detecting obstacles on the footpath that are at knee level. Guide dogs, on the other hand, have more advantages

than white canes, but they also have disadvantages, such as high costs associated with raising, training, and caring for the animal [9]. Furthermore, the use of computer vision and deep learning in assistive technologies for P-VI/blindness has grown in popularity among researchers in recent years. These technologies have created game-changing opportunities for the development of more effective and useful assistive tools for them.

It is important to mention that this thesis has an industrial approach, targeting not only the state-of-the-art challenges but also considering the commercial aspects of the development of an assistive solution. A part of this work was undertaken in a company (GOFORE Spain S.L.) which co-funded and facilitated the process of research and development. They provided the facilities to implement different prototypes of the solution which, ultimately, led to securing some funds for the development of the industrialized smartphone version by being one of the final selected start up ideas of EIT Digital Venture Program 2023.

1.2 Motivation and objectives

This research work will explore how the P-VI/blindness perceive the world around them in indoor settings, how the current solutions provide assistance regarding indoor scene understanding, locating their needed objects, and will identify the main existing issues. This research is focused on indoor scene understanding for several reasons. Both outdoor and indoor scene understanding and navigation have their challenges for the P-VI/blind. However, indoor environments present unique challenges that set them apart from outdoor environments. Outdoor navigation typically allows for the use of environmental cues (e.g., tactile paving, traffic light sounds, etc.) and the white cane, a popular mobility aid that helps in detecting obstacles in the path [10]. Many of these landmarks, however, become less effective or completely inaccessible when individuals with a visual impairment navigate through indoor spaces such as public buildings, shopping malls or any unfamiliar indoor environment. Furthermore, indoor environments usually have more complicated architectural designs, making the navigation and scene understanding tasks more difficult for them [11]. The navigation problem is exacerbated by the fact that assistive technologies that use Global Positioning System (GPS), which are effective outdoors, cannot be used indoors due to lack or imprecision of GPS signals [12]. This implies that a new set of solutions and technologies for understanding indoor scenes must be developed.

This work is a step in that direction, with the goal of delving deeper into the unique complexities of indoor environments and exploring potential solutions for better indoor scene understanding.

Moreover, this work will explore technologies that are potentially useful for our purpose, focusing mainly on computer vision technologies which have improved greatly in recent years and currently represent the most promising option. Nowadays, it is possible to recognize objects using deep learning techniques, more quickly and accurately than ever before. However, while it is currently possible to identify the objects that appear in a scene with a considerable accuracy, it is more difficult to obtain an accurate location of the objects in the three-dimensional space. Measuring the exact distance of an object/obstacle to the user is one of the challenges of creating assistive solutions for blind people. This is because the 3D location of an object in the real world often needs to be obtained from a 2D image taken by a monocular camera. Ultrasonic sensors, point clouds, RGB-D images (taken by stereo vision cameras), or mathematical estimations are solutions proposed to address this problem. Identifying the most appropriate technologies for the intended use cases of a cost-effective specific assistive solution is a challenging task.

Another big challenge is providing a good user experience. To accomplish this, it will be necessary to explore how people with disabilities construct spatial mental models (and the differences between individuals).

According to the points discussed, the main objective of this thesis is:

> Developing a framework for the design of cost-effective assistive solutions for P-VI/blindness, aimed at indoor scene understanding, that helps designers to:
>
> - Understand the needs and characteristics of the target users, to develop more suitable solutions, by providing a list of the possible use cases and the functional and non-functional requirements to be considered.
> - Choose appropriate technologies (hardware & software) for developing such solutions, according to the proposed technology selection guidelines which include different factors such as effectiveness, acquisition cost, computational resources and their limitations.
> - Consider the user experience when designing the interaction with the assistive solution based on the functional and non-functional requirements.

- Design the solution according to a general reference architecture for assistive solutions for scene understanding in indoor environments.

As a secondary objective, the framework will be applied for the development of a cost-effective solution that enables P-VI/blindness to:

- Understand their surroundings in indoor environments.
- Locate objects in their surroundings.

Since this work is an industrial doctorate, two different approaches were considered. The first approach (academic research) was focused on the analysis and evaluation of different state-of-the-art technologies that could be used for the development of the intended solution with a focus on delimiting the current technical possibilities and barriers, and subsequently, the proposal of a framework that would provide designers with tools to overcome the main limitations and challenges found in the state of the art. The second approach (industrialized) was focused on developing a cost-efficient minimum viable product as a commercial solution that could be accepted by the target end users. These two approaches were followed in parallel, leading to the development of two distinct solutions. The first (AssistDiv) is a wearable laptop version that was utilized to assess and analyze the possibilities of various state-of-the-art technologies. Users wearing blindfolds were asked to test and validate our prototype. The second one (V-ASSISTANT) is a minimum viable product for smartphones that was developed to explore the potential for industrialization. It is important to note that the proposed framework was used in the development of both solutions.

2 State of the art

During the past years, various smart assistive technologies for P-VI/blindness have been developed. To fully understand the current state-of-the-art of such solutions, first the existing literature review papers were considered and then, through a systematic mapping review [13] the latest advancements of this rapidly evolving field were investigated. The review highlighted the main challenges of the field and compared different approaches of computer vision-based assistive solutions in recent years. In the following sections, the process of conducting the systematic mapping is explained and then different aspects of the state-of-the-art assistive technologies are discussed.

2.1 Systematic mapping study

The method used for our systematic mapping study (SMS) is the one proposed by [14]. A systematic mapping study's goal is to obtain a detailed overview of a research subject, provide a review of existing literature, identify research gaps, and gather evidence for possible research directions [15]. Consequently, the goal of our research was to gather relevant publications that were related to computer vision-driven scene understanding for P-VI/blindness, and present an overview of the status quo highlighting the research gaps.

A number of research questions were defined to specify the objective of the SMS as follows:

1. What are the current computer vision solutions for scene understanding?
2. How are computer vision methods used to assist blind users with their daily activities?
3. How have proposed solutions been evaluated?

After defining the research questions, the process of collecting papers began.

The major terms used for performing the paper search were "Computer vision," "Visual Impairment" and "Accessibility". By combining these three terms and using similar keywords, the initial set of papers were collected. The list of keywords used in the search is listed in Table 1. The scope of the search considered the papers that were published after January of 2017 in journals, academic conferences, workshops, and academic books. Web pages, non-academic publications and patents were excluded from the search scope. The search strategy is outlined in Table 2. The reason for choosing the papers that have been published after 2017 was that

computer vision methods for object detection have improved drastically since that year and brought significant improvements in the scene understanding research for the P-VI/blindness.

Major terms	Alternative terms
Visual impairment	(Blind people OR visual* impair*) AND
Computer vision	(Computer vision OR visual computing OR object recognition OR image processing OR computational perception) AND
Accessibility	(Accessibility OR scene understanding OR spatial information OR mobility OR assist*)

Table 1 - Search strings

Academic databases	Google Scholar IEEE Xplore ACM Digital Library Scopus
Target Items	Conference papers Workshop papers Journal papers Excerpts of academic books
Search applied to	Title Abstract Keywords
Language	English written papers
Publication period	From January 2017 up to June 2021

Table 2 - Search strategy

According to the research questions, several categories were defined to analyze and compare existing solutions. The three major categories are “Scene Understanding”, “Assistance Services” and “Evaluation”. “Scene Understanding” is related to the first research question and is focused on the level of the perception that the system has from the surrounding environment of the user. The “Assistance Services” category is related to the second research question and defines how the understanding of the outside world by the system is going to assist P-VI/blindness. Finally, the “Evaluation” category collects information about the way in which the solution was evaluated. Each of the topics will be covered in depth in the next sections.

Figure 1 depicts an overview of the selection procedure. The initial number of search results retrieved from all databases was very large (approximately 27,000 for papers published after 2017). To select appropriate papers for the research, the exclusion criteria were applied to the first 50 papers based on their relevance and publication date (most recent). If more than 10 of these 50 papers were relevant to the topic, the subsequent 50 papers were also analyzed. This procedure was repeated until fewer than 10 relevant articles were discovered. The abstracts, introductions, and conclusions of the papers were regarded as the primary source when applying the exclusion criteria (Table 3). In some instances, additional sections of the papers were read to improve comprehension. After removing duplicates (992 results) and applying exclusion criteria, 180 papers were left for full text reading, of which 105 were useful for the review. We excluded a large number of papers because our focus was on those that provided a practical and tangible solution. This signifies that we disregarded papers describing frameworks or solutions lacking implementation (e.g., prototype, proof of concept (POC), or simulation).

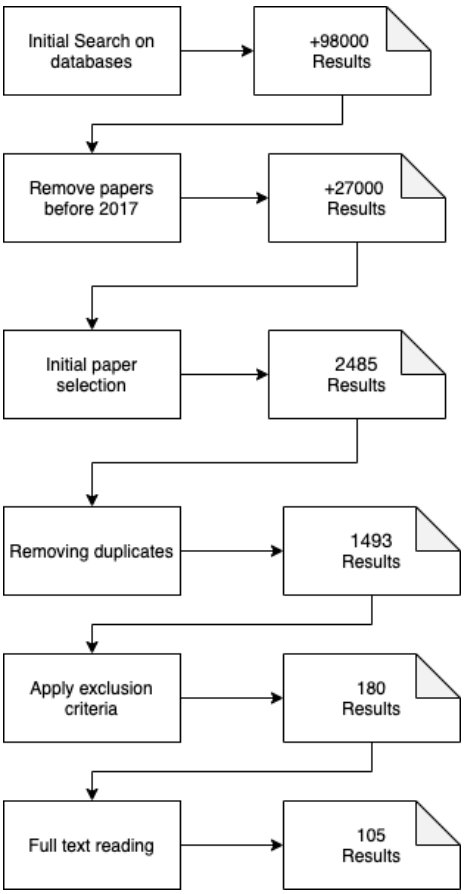

Figure 1 - Paper selection process

Exclusion criteria
1. Papers that are not relevant to the primary research goals. This is obtained after reading the abstract, introduction and conclusion of the papers.
2. Papers published in non-academic web pages, personal blogs**,** or patents
3. Papers without providing a solution for obstacle detection and/or object recognition
4. Solutions that lack implementation
5. Not written in English

Table 3 - Exclusion criteria

2.1.1 Mapping results

The distribution of articles published each year is seen in Figure 2. The number of published papers in this field increased substantially in the recent years. The consistent growth in the number of published articles demonstrates that the subject has received more focus in recent years. This is mostly due to advancements in deep learning algorithms, cloud-based computer vision services, and mobile devices.

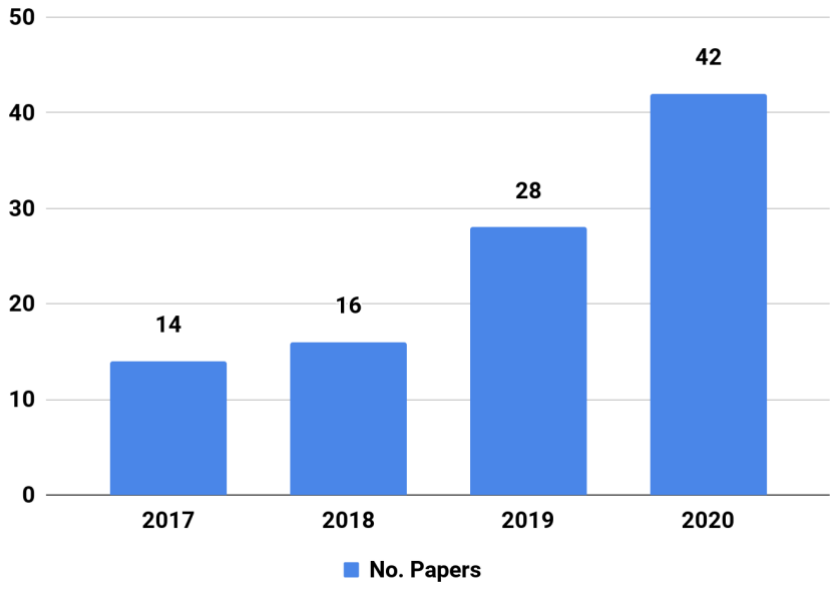

Figure 2 - Paper distribution

2.1.2 Data extraction

A template was created for data extraction for the papers that were selected to be read in full so that the comparison and tracking of the data in the papers would be simpler. To compare existing solutions, a number of categories were defined in accordance with the research questions. "Scene Understanding", "Assistance Services" and "Evaluation" are the three main categories. The phrase "Scene Understanding" relates to the first research question and focuses on how much the system perceives the user's surroundings. The second research question is related to the "Assistance Services" category, which describes how P-VI/blindness will be helped by the system's understanding of the outside world. Finally, information about the method used to evaluate the solution is gathered under the "Evaluation" category. The major categories and subcategories are listed in Table 4.

Scene understanding—RQ1	**Object recognition** **Obstacle detection** **Depth detection** **Algorithms used** **Hardware used**
Assistance services—RQ2	Type of assistance Modality
Evaluation—RQ3	Technical evaluation User testing evaluation

Table 4 - Analysis categories

2.2 Scene understanding

Scene understanding for the P-VI/blindness has some differences with the classical approach. It is very important that the process of analyzing and perceiving the environment by the system occurs in a level that can be beneficial for the P-VI/blindness. For instance, it is crucial that the algorithms have sufficient swiftness and accuracy in detecting/recognizing obstacles and objects to give prompt feedback to the user when it is necessary. Additionally, semantic understanding of the environment by the system and finding the relations between different objects in a scene are important to give a comprehensible description of the environment to a user that has no access to visual cues. After comparing the various research approaches, it was

concluded that previously (since 2000 to early 2010s), most of the solutions were focused on obstacle detection or image enhancement techniques to make it adapted for visually impaired. Image processing was used to make the images perceivable for the visually impaired people. For instance, in [16] [17] researchers used techniques like contrast enhancement, image mapping and magnification. Their aim was to increase the visibility of the important features of an image (e.g., edges). However, these methods had some limitations. For example, the algorithms added too much noise to the image or amplified the contrast of some parts of the image that were not necessary for scene understanding.

Lately, object recognition, which is a more efficient method for scene understanding, has become more popular thanks to technological advancements. The percentage of solutions with object recognition has been increasing in the last years, as shown in Figure 3.

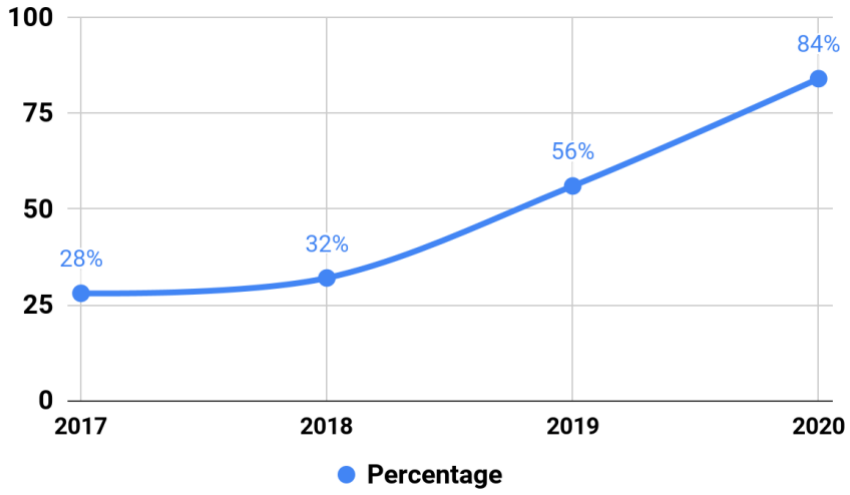

Figure 3 - Frequency of object recognition in the solutions

Mainly, scene understanding for the blind comprises three general aspects. One of the most crucial features of scene understanding is *object recognition*. There are several ways for identifying an item using computer vision, each with pros and cons. Another essential component that must be added in order to notify people and avoid barriers in the surroundings is *obstacle detection* (Figure 4). In this category, we examine each solution's strategy for detecting impediments, whether using sensors or cameras. Finally, *depth detection* is one of the most complicated components of developing assistive solutions. Estimating the 3D position of items in the actual environment in real time using computer vision is still a difficult problem.

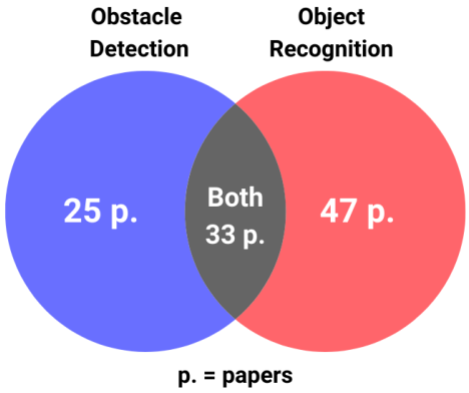

Figure 4 - Solution approaches during recent years

Additionally, other tasks such as *scene recognition* (detecting the type of scene e.g. bathroom, bedroom, kitchen, etc.), *text detection* (detecting texts located in any part of the scene), and *color detection* are a few of the other features that are included in some assistive solutions which will be discussed in the following sections.

2.2.1 Object recognition

Object recognition tasks such as object detection, semantic segmentation, and instance segmentation are all closely related. Object detection identifies the presence and location of objects in an image, instance segmentation identifies the location and boundaries of individual instances of objects, and object segmentation identifies the regions of an image belonging to different semantic classes [18] [19]. All three objectives require learning picture characteristics and applying those features to additional tasks such as image captioning, object identification, and image annotation. These techniques are extremely effective in the field of assistive technology for scene understanding. Next it is described how each of them is applied in various solutions.

2.2.1.1 Object detection

During the past years, the use of deep neural networks (DNN) for object detection, especially the latest convolutional neural network (CNN) models such as ResNet [20] or GoogLeNets [21], has considerably extended the potential of computer vision for developing assistive solutions for the VI/blind. They have a notably superior performance that makes real-time object recognition more achievable in comparison with shallow networks such as AlexNet [22]. In addition, CNNs can learn high level semantic features from the input data automatically,

optimize multiple tasks simultaneously such as bounding box regression and image classification, and solve some of the classical challenges of computer vision [23].

There are different ways to implement an object recognition solution. One common approach is to execute the process of recognition remotely on cloud services like Google Cloud Vision, Microsoft Azure Computer Vision, Amazon Rekognition [24], etc. These services are already trained with huge datasets that enable improved performance. In [25] a mobile system was proposed that uses Google Cloud Vision to recognize objects, texts and faces. There are also solutions that provide their own cloud computing algorithms for image processing. For instance, inn [26] a remote object detector is developed using an improved version of ResNet [20] network. These services mostly have a Representational State Transfer (REST) Application Programming Interface (API) to handle communication between the client and the server. Companies that provide these services usually calculate the costs based on the number of requests sent to the server by the client.

Local image processing is another approach used in many solutions which undertake the computations related to the object recognition on the client side. Nevertheless, this approach is usually confined to a limited number of objects due to hardware limitations. In [27] an android *app* was made for VI people that locally recognizes objects. They used MobileNets [28] for object recognition which is a neural network for mobile and embedded vision applications. Single Shot Detector (SSD) [29] with MobileNets [28] architecture is another popular algorithm which is used in a considerable number of solutions for object recognition and can bring fast and efficient results. SSD can detect multiple objects in an image by taking a single shot. Visual Geometry Group (VGG16) [30], which is a convolutional neural network model, is the base network of the SSD algorithm, followed by a multi-scale feature layer for object category and bounding box predictions. SSD generates anchor boxes in various sizes and predicts objects based on their size. Larger objects are detected by deeper network features and smaller ones are detected by the shallower networks. The inference time of the SSD512 [29] model is 22 milliseconds with about 76.8% Minimum Average Precision (mAP) on Pascal VOC2007 [31] dataset of images, which shows its competence and swiftness in object detection [29]. YOLO [32] is a CNN-based object detection technique and uses Darknet [33] which is an open source network framework written in C and CUDA. YOLO divides an image into SxS grids and generates B bounding boxes for each of them. Afterward, it predicts the probability of classes for objects and their corresponding bounding boxes. In [32], YOLO

VGG-16 had an inference time of 47 milliseconds with 66.4% mAP on Pascal VOC2007 dataset of images, which is close to the SSD performance. Different versions (YOLO v3 (Tiny) [34], YOLO v2 [35],YOLO 9000 [36]) were used in different researches. The accuracy and number of objects that could be detected varies for each version. For instance, TinyYolo is made for mobile devices and is able to recognize a lower number of objects compared with the other versions. Moreover, in [37], [38] Stixel-World [39] was applied, which is a method that is mostly used in autonomous cars, to help P-VI/blindness navigate in an environment. Stixels algorithm provides environmental awareness based on the depth images provided by an RGB-D camera. RGB-D images are captured using cameras that work in association with sensors for distance estimation. Stixels segments objects in the image in front of the user in vertical regions according to their depth and disparity in the environment. Afterward, using object recognition techniques, Stixels semantically categorizes objects in the scene.

In another study [40], object detection algorithms, such as various versions of YOLO, SSD and R-CNN [41] were compared. There is always a tradeoff between precision and the speed of the algorithms. In their comparison YOLOv7 outperformed the other existing object detection algorithms in terms of precision and speed. However, it is important to note that the performance of these object detection algorithms varies depending on the environment lighting conditions, image quality and the objects' visibility and positioning in the image [42], [43]. Therefore, it is important to choose a model that matches the use case of the assistive solution.

YOLO and SSD are very popular because they are achieving a balance between accuracy and speed. Figure 5 shows the distribution of object recognition algorithms in the reviewed solutions. As it is shown in the figure, YOLO is the most popular solution for object recognition. The remaining “Other Neural Networks” in Figure 5 are the algorithms which are used in the solutions are various local object recognition methods that are using Inception-v3 [44], stochastic gradient descent (SGD) algorithms on Keras [45], [46], OpenCV [47] functions or Computer Vision System Toolbox of MATLAB [48], [49], to name a few. Appendix 1 contains the object detection methods used in the reviewed solutions. It is important to mention that the quality of the input images sent to these algorithms can affect their performance noticeably. For example, images taken at night or in low light conditions can have a high noise level or distortion that can reduce the accuracy of object detection algorithms. To overcome this problem, there are different methods. For instance, [50], [51], [52] used a Gaussian filter

that blurs the image in order to remove the noise and unnecessary details in the images before sending them to the algorithm.

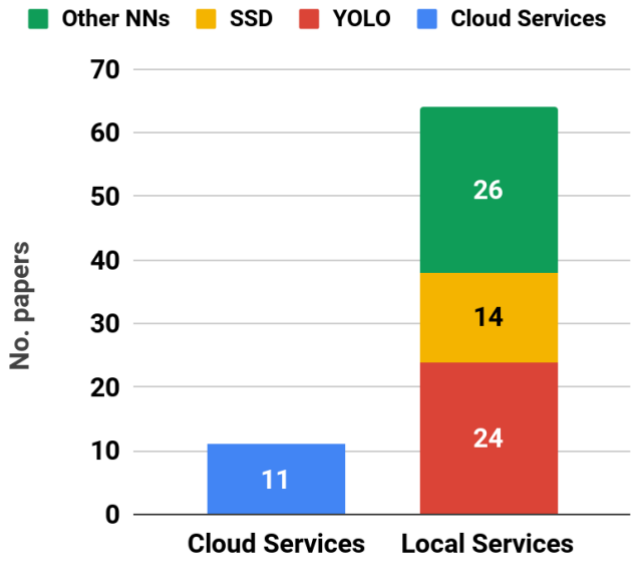

Figure 5 - Distribution of cloud and local object recognition services

2.2.1.2 Semantic and instance segmentation

Another approach to detecting objects in an image is employed in instance segmentation and semantic segmentation. Semantic segmentation involves classifying every pixel in an image into a predefined category [19], effectively providing a comprehensive understanding of the scene by labeling regions with categories such as "road," "sky," or "building." However, it does not differentiate between separate instances of the same object class. Instance segmentation, on the other hand, takes this a step further by not only identifying the class of each pixel but also distinguishing between different instances of the same class [18]. This means that in an image with multiple cars, instance segmentation will recognize and label each car separately.

In contrast to object detection, which locates objects within an image and generates bounding boxes around them, instance and semantic segmentation takes object boundaries a step further by defining them at the pixel level. This increased accuracy makes it possible to perceive a situation more thoroughly, which could be very useful for assistive technology for P-VI/blindness.

Instance segmentation can be roughly divided into segmentation-based methods and detection-based methods. Segmentation-based methods predict pixel-level categories and then aggregate pixels belonging to the same class to achieve instance segmentation. Detection-based methods approach the task by first detecting objects and then refining the detections to produce segmentation masks for each object instance. Detection-based methods trace back to models like Mask R-CNN [53]. Mask R-CNN, in particular, is a popular method that extends Faster R-CNN [54] by adding a branch for predicting segmentation masks, making it effective for instance segmentation tasks. Additionally, recent methods involve hybrid models that incorporate multi-scale convolution and attention mechanisms to enhance the detection of small resolution objects [55].

2.2.2 Object/obstacle location

The process of detecting the location of objects is divided into two tasks. Detecting the object in the scene and finding its location. For the first task, the object detection algorithms that can be used have already been brought up for discussion in the previous section. It is important to note that the same techniques could be used for obstacle detection. Nevertheless, the process of determining the three-dimensional position of the object/obstalce in the surrounding environment is a challenging task. For detecting obstacles in the environment, there are usually two different approaches: distance sensors (e.g., ultrasonic, LiDAR and infrared (IR) triangulation) or camera-based (e.g., monocular or RGB-D cameras) techniques. In some cases, the combination of both techniques is used for better accuracy. Computing the exact distance of an object/obstacle from the user is one of the complications of creating camera-based assistive solutions for blind people. This is because the 3D location of an object in the real world is often inferred from a 2D image taken by a monocular camera. Ultrasonic sensors, point clouds, RGB-D images (taken by stereo vision cameras) or mathematical estimations are the solutions that have been proposed for tackling this problem. In this section, different approaches are discussed.

2.2.2.1 Depth Cameras

Depth cameras are used to provide information about the distance between objects in a scene. Different cameras have been developed to do this task. Various types of depth cameras are described in this section. Each type of has its own benefits, which can vary depending on

the processing power, computational time, range limitation, cost and the environment in which it is employed (e.g. lighting conditions, environment size, colors etc.).

2.2.2.2 Stereo vision cameras

A stereo camera relies on binocular vision, the same principle upon which the human eye operates. Stereo disparity is used by human binocular vision to determine the depth of an object [56]. Stereo disparity is the method of determining the distance to an object by using the difference in the location of the object as seen by two distinct sensors or cameras or eyes in the case of humans. Researchers have worked on this method for decades and developed numerous systems [57]. Nevertheless, stereo vision cameras have certain limitations. The distance between the two cameras imposes fundamental limitations on stereo vision. Specifically, depth estimates tend to be imprecise when large distances are considered, as even very small angle estimation errors result in large distance inaccurate estimations. Moreover, stereo vision tends to fail in texture-less regions of images where pointers for detecting differences in images taken by the lenses cannot be reliably located.

Some cameras (e.g. Inteal RealSense D455) include a projector besides the stereo camera. This method employs the projection to either simplify the correspondence matching between the camera images (Infrared (IR) projector that projects a dot pattern). This approach has a significant advantage in low-light environments or when the texture is not very distinctive.

2.2.2.3 Structured Light Cameras

A structured light-based depth sensing camera projects light patterns (typically stripes or speckled dots) onto the target object using a laser or LED light source. The depth is determined by analyzing the observed distortions in them. (e.g. first version of Kinect or the Intel RealSense SR305 [58]). These cameras provide high - precision depth information at a relatively high resolution and function well in low-light conditions. However, the detection range of SR305 and similar cameras are not promising in long range. In [59] findings demonstrated that the camera works the best at short range and significantly worsens in distances greater than 55 cm. In addition, some structured light cameras require more computational power and processing time to operate since they require multiple projections in comparison with stereo vision cameras. Such cameras do not work well on highly reflective, transparent, very dark surfaces or in outdoor environments where the lighting is intense [60].

2.2.2.4 Time-of-Flight (ToF) Cameras

Time-of-Flight (ToF) cameras (e.g. Kinect Azure [61]) acquire depth data by calculating the time required for a light ray to travel from a light source to an object and return to the sensor. In the case of Kinect Azure, the maximum distance range for depth detection is up to 5.46m [62]. Other types of ToF cameras such as laser-based cameras (LiDAR) are more suitable for longer range depth detection. For instance, Intel RealSense LiDAR Camera L515 can estimate the depth up to 9 meters [63]. LiDARs are commonly utilized for 3D measurements in outdoor environments. The primary benefits of LiDAR sensors are their high resolution, precision, low-light performance, and speed. However, LiDARs are costly and not energy-efficient devices, making them inappropriate for consumer products [64].

2.2.2.5 Monocular-depth estimation methods

Besides the discussed methods, there are some other approaches such as deep learning-based methods that became popular during the recent years. It has been demonstrated by the current state of the art that monocular depth estimation methods could be a viable solution to various depth-related challenges. In [65] a depth prediction network was built that provides a depth map from a single RGB image. Their predictions work with images taken with different camera models. Various kinds of Neural Networks (NNs), like CNNs [66] and RNNs [67], have been implemented showing the effectiveness of monocular depth estimation. In [68] a CNN network was used for calculating the distance of the obstacles. Their method works more accurate than some devices like Kinect, according to their comparisons.

These methods require a relatively small number of operations and less complications compared to more complex depth measurement techniques that involve multi-camera or multi-sensor setups. In contrast to stereo vision or other sensor-based depth measurement methods, monocular depth estimation does not require alignment or calibration. Moreover, the ability of monocular depth estimation to operate with a single camera viewpoint allows for greater flexibility in deploying computer vision systems across different environments and platforms. Accurate monocular depth estimation methods can contribute significantly to the comprehension of 3D scene geometry and 3D reconstruction, especially in economical applications [69]. However, the performance of these methods heavily depends on the data that the models were trained on.

The generalization ability and reliability of a deep learning model are largely determined by the quality of the datasets. For instance, a deep learning monocular depth detection model that was trained in daylight might not have a good performance at night time [70]. To enhance the accuracy of depth estimation, it is necessary to collect more data of higher quality and from a wider variety of scenes. Nevertheless, these existing datasets used for depth estimation are limited, and creating a new dataset is both time-consuming and costly [71].

In [72] a mathematical method was used which is proposed in [73] for object distance estimation. In [72] they claimed that this method has a high accuracy in distance estimation and can detect an object's distance up to more than 10 meters. The solution in [74] is based on another mathematical method that uses depth images and fuzzy control logic for the approximate measurement of obstacles' distance. This solution divides the frame into three parts (right, left and center), categorizes the location of obstacles in three different categories and provides audio feedback for the user according to them. If the user faces any obstacles, the system makes decisions in order to avoid them based on 18 different fuzzy navigation rules that depend on the location and distance of the obstacles.

2.2.2.6 Distance sensor-based techniques

Using sensors has been a more common approach for obstacle detection in comparison with camera-based techniques. Among the different kinds of sensors, ultrasonic is very popular. This is because of their accuracy, low cost, low power consumption and ease of use [52]. Ultrasonic sensors have a transmitter that generates sound waves with a frequency that is too high for human ears to hear. Then, the receiver of the sensor waits for the rebound of the sound and based on that, the distance with the obstacle will be calculated. These sensors are better at detecting transparent objects compared to light-based sensors or radars. They can be mounted on a walking stick or other mobility aid that produces a beeping sound or vibration when an obstacle is detected. The advantages of ultrasonic sensors are their capability of detecting dark surfaces [13], working in dim lighting, detecting transparent objects and low power consumption. Many solutions such as [52], [75], [76] were developed utilizing these sensors. However, there are a few drawbacks, including a short detection range (about 2 meters) [77] and poor accuracy when detecting soft or curved objects.

As an example, in [25] ultrasonic sensors were used for the detection of knee level and low-lying obstacles. One of the drawbacks of using sensors is that they cover a short range and can only detect close obstacles, which makes them more suitable for indoor environments.

Infrared sensors, on the other hand, utilize infrared radiation to detect the presence of objects in the environment and can also be used to identify obstacles. They are effective at detecting soft objects and are better compared to ultrasonic sensors at detecting the edges of objects. However, the detection range of infrared sensors (from around 20 cm to 150 cm [76]) is quite limited.

2.2.2.7 Combined distance sensor and camera-based techniques

Lately some researchers have combined these two approaches. In [52] ultrasonic sensors and RGB-D cameras were used for obstacle detection. Their electronic travel aid (ETA) processes the data received from an RGB-D camera using a Raspberry pi 3 B+ which has ultrasonic sensors attached to it for distance estimation. The combination of these two approaches provides a more accurate obstacle detection. Appendix 2, contains the detail of the different obstacle detection techniques that were used in the reviewed solutions.

Additionally, in a couple of studies [78], [79] the Simultaneous localization and mapping (SLAM) method is used for navigation and obstacle avoidance. SLAM [80] can determine the position and orientation of the sensor relative to the surrounding environment while also mapping the environment. Visual SLAM [81] is a type of SLAM that uses camera image input to perform SLAM positioning and mapping in unknown environments. Researchers in [82] proposed a blind guidance system using both tactile and auditory feedback by combining ORB-SLAM [83] and YOLO. The system had the capability to identify the precise category of obstacles and provide their location through real-time voice messages, and to plan paths avoiding obstacles. To avoid obstacles, an obstacle avoidance algorithm based on a depth camera was implemented. Simultaneously, an algorithm for converting the sparse point cloud generated by ORB-SLAM into a dense navigation map was developed. In addition, image recognition based on the YOLO algorithm was used in real-time to detect the obstacle target. SLAM methods for assistive solutions are still recent, but they promise to be an influence in the future of navigation and obstacle avoidance.

2.2.3 Scene recognition

When sighted people look at an image of an environment, they can comprehend different aspects of it, such as what is occurring, who is engaged, what is shown in the image, and how the various elements are related to one another. Through this information, it is possible to identify the environment and situation they are in. However, people with a visual impairment face difficulties in this regard. In order to address this issue, incorporating a scene recognition feature in an assistive app could be extremely beneficial [84].

Scene recognition using computer vision is an attempt to recognize the image's content, the items present, their locations, and the semantic links between them [85].

Early scene recognition methods (in the beginning of 2010s) were mainly based on global attribute descriptors which are composed of a few basic visual characteristics to simulate human perception [84]. Global attribute descriptors offer a high inference speed since they can be generated using some pre-defined numerical computations without the need for training. However, they are only able to collect a small number of basic visual cues, which restricts their capacity and results in low accuracy [86]. Subsequently, to boost the performance of recognition, patch feature encoding was introduced. More resilient than global attribute descriptors, algorithms based on patch feature encoding can handle complexities such as cluttered backgrounds and object deformation within a specific region. Nonetheless, neural networks, particularly deep learning models [87], consist of multiple layers of interconnected neurons, each designed to progressively extract higher-level features from input images. As these networks become deeper and more complex, they can learn highly detailed patterns. However, this complexity also leads to increased computational requirements, especially during the inference phase, where the network processes new inputs to make predictions. Each additional layer or image patch adds to the computation load, thereby slowing down the system. In general, the techniques based on patch feature encoding can be used in some particular use cases when computational resources and scene categories are constrained, and reaction time is more important than recognition accuracy [86].

Later (since 2017), with the advancements in deep learning and convolutional neural networks (CNN), more advanced methods such as the ones in [88], [89], [90] was presented with a high classification performance. For scene recognition using CNNs, features from intermediate and high layers are more useful because they represent more sophisticated concepts (e.g. parts and

objects) than those from low layers (e.g., edges and textures). Low layers usually contain repetitive or correlated information [86].

During the past years, several “Hybrid” models that try to combine the power of feature encoding and end-to-end networks have evolved. Patch features and global features from the end-to-end network's multi-stage outputs are used for feature encoding in order to create image representation. For example, FOSNet [90] is suggested to merge object and scene information in an end-to-end CNN framework which assumes that the neighboring patches of a single image belong to the same scene class.

Additionally, a significant number of annotated datasets have contributed to the process of training a model with high accuracy. There are various datasets such as Scene-15 [91], UIUC Sports [92], MIT Indoor-67 [93], SUN-397 [94], Places, that can be used for the training of the scene recognition algorithms. Among the mentioned datasets, MIT Indoor-67 [93] is focused on indoor environments making it a more suitable dataset for our goals.

Since assistive devices are mostly portable and have a limited computational power, it is possible to train the current scene recognition models using lite models which are for mobile devices. A scene recognition solution is proposed in [84] for the people with VI/blindness that is based on a EfficientNet [95] lite model. Their solution can smoothly run on a smartphone and detect 15 different scene categories. With some accuracy compromised but improved inference speed, several of the current deep learning models, including MobileNet, YOLO [32], and Inception [30], can be trained for scene recognition on mobile devices.

2.2.4 Text detection

Scene understanding also involves being able to read the texts that are present around us. Optical character recognition (OCR) is a software technique that recognizes and extracts text from images and documents. OCR can be used to help people with blindness read by recognizing text present in a scene, that can then be read aloud by a text-to-speech software. There are some current solutions such as SeeingAI [96] that include this feature. The 'Short Text' mode of SeeingAI is a straightforward, potent, and quick OCR application. It has a simple user interface, which contributes to its appeal. Once the application is activated, it simply reads aloud any text it detects in images continuously captured by the phone's camera. There are some open-source projects such as Tesseract [97] by Google with a fairly acceptable accuracy [98] in detecting text. However, if the user is moving around, automatic detection and reading

of scene text may produce poor results. Two major issues contribute to this being a difficult problem. First, images captured by a moving camera can be blurry, complicating OCR's work. Second, when exploring a scene in search of a sign with a camera with a limited field of view, multiple pictures in different directions are required. To provide timely feedback, the system must be able to process these images quickly. However, the designer may have to sacrifice speed for image resolution, which is required for distant reading [99].

2.2.5 Color detection

Color detection plays a vital role in enhancing the independence and quality of life for individuals who are blind. By providing the ability to recognize and differentiate colors, this feature opens up new possibilities for engagement with the world, from selecting outfits to identifying objects and navigating spaces. It allows for a more inclusive experience, ensuring that those with visual impairments can fully participate in daily activities that are often dependent on color recognition.

Instance segmentation technique, which was explained previously in object detection methods, could be utilized for detecting the color of an object too. This approach can detect the color of objects by getting the HSV (hue, saturation and value) and RGB (red, green and blue) values of a pixel and turning it into natural language [100], [101]. However, color detection becomes a challenge when the object has textures, patterns, and a wide range of colors. In such cases, describing the object's color in natural language remains an open problem. As a result, existing solutions are mainly limited to describing plain colors.

2.3 Commercial solutions

There are some commercial solutions in the market for the P-VI/blindness that try to address some of the mentioned challenges. For instance, mobile apps like Envision [102] provide features like scene description, text reading, barcode scanning, color detection, object finding and so on. They also provide smart glasses that have the same features in a wearable medium. Although their app is free to use, their glasses are costly (starting at 1900 $). There are other similar apps like SeeingAI [96] and Lookout [103] by Microsoft and Google respectively. Other apps, such as Aira [104] and Be My Eyes [105], use the smartphone camera of the person with VI/blindness to deliver actual human assistance for various tasks such as scene description or text reading. Aira provides trained human interpreters for that purpose and Be My Eyes is a crowdsourcing approach that might have a lower quality of assistance, but it supports more

languages, and it offers more interpreters. However, despite their helpfulness, some privacy concerns arise for some users as a real person connects to the user's device [106].

Be My Eyes has recently added a feature to the app called BeMyAI [105] that uses the power of multimodal large language models (MLLM) which supplement text-based knowledge with additional data modalities such as visual data [107]. The user can take a photo from the scene and the app will provide an extensive description about the photo which includes type of objects, colors, texts, textures, and the positions of the objects in the frame. Then the user can ask for further information about the photo taken via chat. It provides more detailed information in comparison with other scene description solutions in the market, but the accuracy of information is something that needs to be evaluated as these models can hallucinate in some cases [107].

2.4 Assistance services

It is crucial to consider how the information regarding the environment obtained by sensors, cameras and algorithms is transferred to the user. The assistance provided by these solutions should be swift, accurate and easily understandable for P-VI/blindness. The assistive solutions reviewed help users in different tasks relating to their daily life. In [108] a list of user requirements is provided after a thorough literature review on studies about requirements elicitation for assistive solutions for VI/blind people. The requirements considered in our SMS according to their study are listed in Appendix 3. Obtaining information from the surroundings is considered as one of the most important requirements for the assistive solutions for P-VI/blindness, because it helps them to create a more accurate mental map of the environment [109]. Therefore, the capability of computer vision technologies to analyze and understand a scene has been considered as the core of our mapping study.

The modality of existing assistive solutions is mostly based on audio and tactile feedback. Binaural audio for scene description is used in [2], [3], [110]. This brings a sense of audio-based augmented reality to the user. Users can hear and feel the approximate 3D location of the object/obstacle based on the audio they hear using normal/bone conduction headphones. In these solutions, the device/*app* names all the objects in the scene or a specific object that the user is looking for. Moreover, some solutions provide vibrotactile feedback. In [101] two servo motors were used for vibrating feedback so that when there is an obstacle on the left, the left side of the user vibrates and when there is an obstacle on the right, the right side vibrates. In

another research [111], the user is warned about the obstacles with an audio beep. The pitch of the sound changes based on the size and proximity of the object. There are solutions that carry out the scene description for a specific purpose. These solutions might be limited to certain tasks, but the overall performance is better because the scope is constrained. In [45] stop lights and crosswalks are detected to help users with crossing streets.

Many solutions provide navigation assistance for users. For instance, [112] provides a navigation service that guides the user through a pre-scanned environment. A virtual assistant repeatedly states "follow me" and, based on the intensity and the direction of the sound in the headset, the user navigates through the environment.

Besides the scene understanding feature, emergency calls are provided in some solutions which can be very useful for the users. In [113] a speech recognition module was implemented that can get orders from users to make an emergency call to predefined users or the closest emergency center based on the location of the user. GPS was used for tracking the live location of the user. In [114], [115], [25] a similar approach was undertaken by placing a call button on the assistance device.

In [116] a text recognition module was proposed in their solution that reads the text that is in front of the user. They used Google Vision API [117] for this purpose. The same functionality has been provided in [118] using Tesseract OCR library [119] and in [120] using Microsoft's Computer Vision API [121]. However, the usage of this feature can be confusing for the user since it is not possible for P-VI/blindness to distinguish the exact position of a text in the real environment.

In [122] a face detection module was included in their solution besides object recognition. They used Multi-task Cascaded Neural Networking (MTCNN) that can detect faces with a 70–100% accuracy. The kinds of assistance provided in the revised solutions are listed in Appendix 4.

2.4.1 Context of use: the ICF framework

In order to come up with novel and useful solutions for any kind of disability, it is necessary to understand the contexts in which they could be applied and the scope of limitations that a disabled person faces. The International Classification of Functioning, Disability and Health (ICF) framework [123] provides a standard language for defining different kinds of disabilities. In the framework, limitations of disabled people in their activities are divided into different categories. In our research, we tried to assess the usefulness of the proposed solutions to P-

VI/blindness by mapping their contributions with the different tasks in this framework. In the case of scene understanding for P-VI/blindness, the main relevant categories are as follows:

- ***Mobility.*** This is mainly about moving the body and going from one position to another. According to the framework, it includes tasks like "walking and moving," "changing and maintaining body position" and "moving and handling objects."
- ***Self-care.*** This includes tasks like washing oneself, caring for body parts, toileting, eating, dressing, drinking, and looking after one's health.
- ***Domestic life.*** It covers daily tasks related to acquisition of necessities, household tasks, caring for household objects and assisting others.
- ***Interpersonal relations and relationships.*** This category consist of general interpersonal interactions and particular interpersonal relationship challenges that a disabled person may have with other people.

We analyzed the compatibility of the above mentioned ICF categories with the solutions provided by different researchers. Mobility is the most explored category, with research in [112], [124], [125] helping P-VI/blindness with navigating in different environments (indoor/outdoor). They combine various methods like GPS, obstacle detection and object detection to help the P-VI/blindness in tasks like "walking and moving" from one location to another. Moreover, [49] defines two user stories that are compatible with the "Domestic Life" and "Self-care" categories. In the first user story, the blind user receives assistance for detecting the right kind of pasta at home, which helps her with cooking and eating. In the second user story, a man wants to buy a specific kind of biscuits and using the assistive device he can find them in the store. Their solution provides feedback about the differences in objects that have the same tactile appearance.

By comparing the categories of the ICF framework with the kinds of assistance provided in the revised solutions, we noticed that their scope is generally not well defined in terms of the context of use. For instance, in papers like [35], [113], [124] object detection could potentially cover some tasks in "Self-care" or "Domestic life." However, these specific use cases are not mentioned. Furthermore, in [122], [126], [127] face detection is provided, a service that could be related to the "Interpersonal Relations and Relationships" category, but the final purpose is not clear. It appears that the researchers have focused their efforts more in proving the technical feasibility and performance of the solutions than in demonstrating how the solutions can help

P-VI/blindness in their daily life. This fact becomes more evident when we analyze the way in which the solutions have been evaluated.

2.5 Evaluation

Based on our review, the evaluation process of an assistive solution should tackle two main aspects. One is the technical evaluation and validity assessment of the system from a technical point of view, and the other is the testing of the system with the target end users to evaluate the performance and usefulness of the solutions. Despite the fact that technical evaluation is a matter of importance, user testing is equally essential. This is because the ultimate goal of any assistive solution is to be useful for P-VI/blindness. Unfortunately, user testing is neglected in a considerable number of papers. Figure 6 shows the percentage of papers that only undertook the technical evaluation, and the ones that performed both.

2.5.1 Technical evaluation

In the development process of any system, it is crucial to test and measure its performance with objective metrics. In the case of assistive solutions, it is essential to measure the accuracy and efficiency of algorithms in the scene understanding. It is common to evaluate the performance of object detection algorithms using the calculation of precision, recall and minimum average precision (mAP) which is the mean of average precision calculated out of precision and recall metrics. mAP is calculated by using the following formula:

$$mAP = AP/N$$

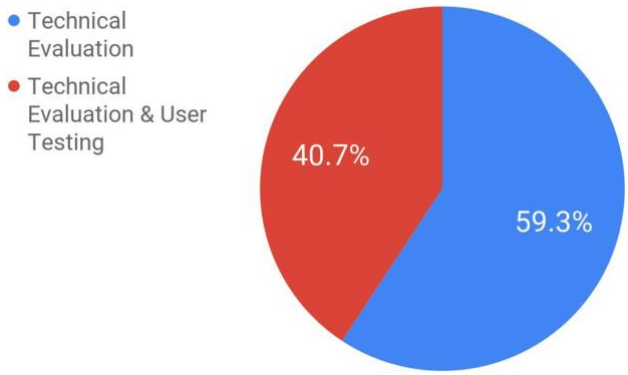

Figure 6 - Evaluation approach

N is the numbers of object category. The formula represents that calculating the average precision (AP) for each category and averaging the AP of all categories.

A model with high precision returns more correctly predicted results than incorrect ones and a high recall means that the model returns most of the relevant results. In other words, precision is a measure for quality, while recall is a measure for quantity. Some papers, such as [51] and [111], measured the accuracy of their approach based on recall and precision. Others, like [45], [125], used mAP to measure the effectiveness and accuracy of the algorithm they used. Additionally, other works [128], [129], [130], [131] tried to compare their solution with other state-of-the-art solutions based on the accuracy, number of detected objects, kind of recognition (object, text, face, obstacle, etc.), average distance, convenience and so on, in order to assess the functionality of their solution and ascertain the competence of their approach. In technical evaluation, researchers usually test their system by simulating a user scenario. For instance, in [38] a prerecorded video was used to test their Stixels model.

2.6 User testing

2.6.1 Participants in user testing

The most important aspect in the evaluation of a solution for P-VI/blindness is to test it with the target users. This is because assistive solutions are for a target group that is different from average users. A person without the disability cannot evaluate the solution properly, given that, due to the variation in sensory input, they do not possess the same mental models of the environment and qualities of embodied experience as people that genuinely have visual impairment. Sadly, this critical point is neglected in a noticeable number of research projects. Up to 27.5% of the papers that were included in our review were testing their solutions with blindfolded/sighted users. Figure 7 summarizes the visual perception status of the testers in evaluations. Nonetheless, there are some works that report testing with P-VI/blindness. In [132] 100 hours of testing for the navigation module of their solution with simple and complex paths was performed. Afterward, they asked the 5 blind testers to fill in questionnaires regarding the comfortability and effectiveness of the system. In another research [111], the solution was tested with 13 participants that were P-VI/blindness. They conducted 15 minutes test sessions and asked users to fill in a Likert-like scale questionnaire. Furthermore, [133] reports a detailed evaluation and found that there is a significant difference between early blind and late blind testers, and that the former group could perform better. The authors also came to the conclusion that vibrotactile cues are less efficient in comparison with auditory cues for detection in the central region of the environment. Additionally, in [134] which presents a navigation robot for

the P-VI/blindness, a satisfactory user study is included. They evaluated their solution with 10 blind participants. The tasks for testing are well designed and explained in the paper. After the testing process, user feedback was obtained about confidence, safety and trust in the solution using questionnaires. The results obtained from questionnaires are subjective and contain personal opinions of the users. In some cases, personal opinions can be considered unreliable when they are obtained from blindfolded users instead of P-VI/blindness. For example, in [37] researchers used the NASA- TLX [135] evaluation questionnaire that analyzes tasks based on Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort and Frustration. They noticed that users reported unexpected low scores on physical demand (which means they needed demand). This is because they were blindfolded and the tasks appeared to be more frustrating for someone with normal vision in comparison with P-VI/blindness.

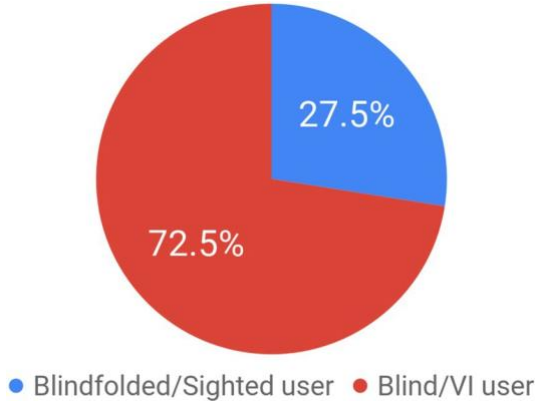

Figure 7 - Evaluation with sighted/P-VI/blindness

2.6.2 User testing methods

The main methods used for testing the solutions were as follows:

- Surveys: Asking a series of questions from the users, usually with Likert-type scales, to obtain their feedback;
- Think-aloud protocol: Users share their opinion about the solution while performing the test tasks;
- Controlled environment testing: Testing the solution in a laboratory environment that was designed by the researchers and observing the user's behavior to detect the advantages and problems of the prototype;

- Field experiments: Testing the solution in real-world settings with the subjects. In this type of testing, users might make unexpected decisions which help to find out the scenarios that were not considered by the researchers;
- Remote usability testing: Users are not directly observed while using the assistive solution. Data are gathered and then later analyzed by the researchers;
- Interviews: Users are interviewed to share their opinion about the experience of using the solution and its pros and cons.

Each of these methods have their drawbacks. According to [136] representative surveys cost money and time, the think- aloud protocol is not accurate because the environment is not natural to the user and the tasks are usually performed in a controlled environment; field experiments may not represent the correct population; remote testing needs additional tools for collecting data; and interviews do not sufficiently cover usability issues. Additionally, controlled environment testing might not consider some factors that exist in the real environment which may affect the user's experience. Testing of the systems should be performed in a real-world scenario to assess if they are usable out of the laboratory's controlled environment. If the usability or performance of the system is not properly tested in the real context of use, it can end up causing negative impacts on the target end user.

Appendix 5, details how user testing was performed in the selected papers. The ones, that are not included, only had performed a technical evaluation.

2.7 Adoption of assistive technologies

The exclusion of VI/blind people in the evaluation process, is an aspect that may be critical for the adoption of the technology. In [137] it is suggested that some of the reasons for abandoning assistive technologies are:

- Neglecting users' opinions in the development process.
- Inefficiency of devices .
- Insufficient training of the user.

Furthermore, in [138] four factors are identified similar to the previous research, related to the abandonment of assistive devices:

- Change in user needs and priorities in time is one of the main factors in device abandonment. For instance, according to [139], some changes in P-VI, like worsening

eye condition due to macular degeneration, can imply a significant change in user needs.

- Not considering users' opinion in planning and decision making during the design and development.
- Lack of easy device procurement.
- Poor device performance.

2.7.1 Privacy issues

Despite the fact that these solutions help users overcome social barriers and give them more independence in life, computer vision models rely on camera input which could threaten the privacy of users and surrounding people. One of the greatest risks is that the collected data get misused, especially in solutions that rely on remote servers for computation instead of the users' devices. Some studies show that there is a trade-off between the provided services and the privacy costs, and that some users are willing to accept the privacy costs in exchange for the service they receive [140]. In [141] a study was conducted about the social acceptance of assistive solutions for the blind from the perspective of both blind users and bystanders. They concluded that a considerable amount of people in the society are still not very comfortable to be exposed to these devices, especially if they include a camera. Their results indicate that a thorough evaluation in a real environment is needed to evaluate the social acceptance of assistive solutions and the needs of people who are exposed to the technology. In [142] a similar point of view was shared. Both sighted and non-sighted users in their study were concerned about their privacy and the accuracy of the information provided by the assistive solution.

2.8 SoA conclusions

Our study suggests that researchers and designers in this field should pay more attention to the P-VI/blindness needs and the real world applicability of the solutions. Many of the papers reviewed in our systematic mapping had insufficient/no data regarding the target context of use for the proposed solutions. Consequently, it was not possible to assess the compatibility level of their solutions with the ICF categories and their specific tasks, such as dressing, eating, drinking, preparing meals, acquisition of necessities and so on. There are solutions that prove to be able to detect persons, objects or obstacles, but their expected benefits in the end users' life are vague. This situation raises important generalizability concerns, given that a solution just tested in a toy or simulated scenario might not have the same effectiveness in different use

cases or scenarios. The cost and effort associated with the adaptation of a specific solution for its application in a different context is generally overlooked.

Moreover, the way that information is represented and delivered to the user by the solutions is crucial. Users should be able to comprehend the information provided by the solution without complications. In most of the solutions, binaural audio, vibrotactile feedback and basic audio instructions were used for this purpose. However, these might not be always the best approaches. For example, binaural audio can approximately indicate the location of an object or an obstacle, but still it is not very accurate. The mentioned challenges may have contributed to the abandonment of existing commercial assistive solutions, leading to the unavailability of them in the market. Paying more attention to the ICF framework, user requirement studies and the UCD methodology would be desirable for developing more usable, useful and better accepted assistive solutions. Besides that, learning from some of the success stories of products like WeWalk [143] that was designed by a visually impaired person known as Kursat Ceylan, who was familiar with the needs of blind people, can lead us to the development of more useful assistive devices.

Finally, given the variety of software and hardware technologies available, it can be difficult to understand why developers choose certain methods and tools over others when developing assistive solutions. The wide range of options, each with their unique features and limitations can introduce uncertainty and complicate the decision-making process for developers seeking to design and develop such solutions. This complexity emphasizes the need for a more structured approach to evaluating and selecting appropriate technologies.

3 Problem statement

In order to attain the primary objective of this thesis which is developing a framework for the design of a cost-effective scene understanding solution that helps P-VI/blindness, and after the analysis of the state-of-the-art, three main challenges regarding the development of such solution clearly stand out, as follows:

- The integration of user needs and requirements during the design and development process, discussed in section 3.1.
- The selection of appropriate technologies (both hardware and software) among various available options, that align with an intended solution, addressed in section 3.2.
- Taking care of the user experience and effectively communicating feedback from the system to users, explored in section 3.3.

The specific scope of these challenges will be outlined in section 3.4.

3.1 User-centered design and development

For the development of any solution, it is crucial to understand the context of use before specifying user requirements, proposing design solutions, and evaluating them against the requirements.

It is especially critical to learn about the characteristics, needs and concerns of users with a P-VI/blindness in order to design a usable assistive solution. In the systematic mapping [13], we gathered and analyzed the solutions that had been developed in this field from 2017 to 2020, and among the gaps that we found were the lack of user-centered design and the absence of sufficient user testing with the target population. In the systematic mapping, around 30% of the solutions had only been tested with sighted/blind-folded users which shows a significant neglect of the target end users (Figure 7).

There is not much assistive technologies research that focuses especially on user understanding and requirements elicitation with users with VI/blindness.

In [144] the effectiveness of various indoor navigation solutions for the people who are blind was evaluated and some of the common requirements for such solutions, such as ***accuracy, less computational load in portable devices, providing audio output, identification of landmarks, and minimizing cognitive load of the user*** were discussed. In their research, they discovered

two major issues: the necessity to assess the applicability of a navigation solution to a variable indoor setting and the crucial importance of thoroughly knowing the specific needs, goals, and skills of the people who are blind or have a visual impairment around the world.

Hersh and Johnson [145], conducted a global survey, and offer a thorough investigation of the attitudes, interests, and needs of individuals with blindness for the development of a robotic guide. Respondents expressed interest in the guide robot doing a variety of tasks, like avoiding obstacles, navigating, reading information, crossing a road, using public transportation, responding to emergencies, and localization. This research also remarked some of the requirements and design issues with the assistive solutions, like the ***device's aesthetic (not attracting much attention), portability, battery life, and ease of use (accessible interface, robustness, user adaptation, etc.).***

It is also important to consider that visual impairment has a spectrum and people can have a variety of visual conditions that can affect the kind of assistance they need. In [146] a survey was conducted to find out how assistive glasses for the P-VI/blindness should be tailored according to distinct visual pathologies. For instance, in the case of face recognition, participants with optical nerve lesion (ONL) expressed low interest, while the ones with retinitis pigmentosa (RP) and glaucoma (GI) only required it for peripheral vision. Additionally, there is a substantial difference between the experiences of users who are totally blind, deaf blind, or with a visual impairment.

Their work demonstrates that such **solutions must be customizable** and that P-VI/blindness should not be regarded as a single population and their individual differences must be recognized.

Other papers highlight the trust issues that some people with blindness have with such solutions. For instance, in [147] which is a study about requirements for tactile mobility systems, users expressed their ***concern about the system running out of power or not having sufficient accuracy***. Some users were worried about ***not being able to interact with the assistive system because of its complexity***. This concern is amplified in instances where there are no bystanders who can assist people with blindness in rediscovering their path. While dogs may be trained to navigate in familiar environments, they are sometimes not able to provide assistance when entering an unfamiliar situation. Failure of devices in such unfamiliar contexts influences their acceptability.

In a separate study, Akter [142] investigated the social acceptability of such solutions. Both sighted and non-sighted users were concerned about ***their privacy and the accuracy of the information the assistive solution provided.***

In conclusion, an assistive solution should be designed according to the user needs, concerns, and their level of visual impairment. Some of the proposed solutions may not be very useful for and accepted by the end user as a result of neglecting a proper analysis of the context of use (the user characteristics, needs and concerns, the tasks to be supported by the assistive solution, and the environment where the solution should operate).

There is a lack of a clear framework for the relevant tasks and environments an assistive solution could address. Most of the current solutions have focused on outdoor navigation, with less research performed for indoor navigation and scene understanding. Both outdoor and indoor scene understanding, and navigation, have their challenges for the P-VI/blind, but our research is focused on indoor scene understanding for several reasons. Indoor environments present unique challenges that set them apart from outdoor environments. Outdoor navigation typically allows for the use of environmental cues (e.g., tactile paving, traffic light sounds, etc.) and the white cane, a popular mobility aid that helps in detecting obstacles in the path [10]. Many of these landmarks, however, become less effective or completely inaccessible when individuals with a visual impairment navigate through indoor spaces such as public buildings, shopping malls or any unfamiliar indoor environment. Furthermore, indoor environments usually have more complicated architectural designs, making the navigation and scene understanding tasks more difficult for them [11]. The navigation problem is exacerbated by the fact that assistive technologies that use GPS, which are effective outdoors, cannot be used indoors due to lack or imprecision of GPS signals [12]. This implies that a new set of solutions and technologies for understanding indoor scenes must be developed.

Our work in an attempt in that direction, with the goal of delving deeper into the unique complexities of indoor environments and exploring potential solutions for better indoor scene understanding. In order to provide designers with tools that help them in the analysis of the context of use, we propose a systematic categorization of the various scenarios that the user might run into for an indoor scene understanding solution. The scenarios are described as use cases in chapter 6. These scenarios are intended to address the unique challenges associated with understanding indoor scenes. They are defined based on our previous systematic analysis of literature [13] as well as a primary user research which focused on several challenges that

people with P-VI/blindness face in their daily lives regarding indoor spatial awareness (Chapter 5).

3.2 Technology selection

It is necessary to assess a number of variables before choosing the best technology to implement a solution for people with a visual impairment/blindness, including the **computational resources required, cost, portability, simplicity, price, speed, battery life, aesthetics, and feedback type** [148][149][150]. Simultaneously, it's crucial to consider the delicate balance between the solution's performance, functionality, and affordability during the selection process to develop a useful solution. The selection of a technology necessitates a comprehensive understanding of these trade-offs. The goal is to optimize the solution's **effectiveness** and **efficiency** without compromising its **accessibility** or **usability**, thereby achieving the best possible outcome for the intended users.

Since one of the objectives of this thesis is to provide a solution which is cost-efficient and has the potential to be commercialized, we decided to only consider the technologies that could be used on the mid-range price devices (not very expensive) without the need of acquiring costly hardware or services.

To address this problem, we propose a method for selecting an appropriate technology according to various aspects such as effectiveness, computational resources, limitations, and acquisition cost. This will include analyzing a variety of approaches that combine different hardware and software to assist developers and researchers in selecting appropriate technologies for their work.

3.3 User experience

As it is discussed in the section 2.4 assistance services, the way information provided by the solution is communicated to the P-VI/blindness and the user interaction is very important. It is crucial to consider how the information regarding the environment obtained by sensors, cameras, and inferred by algorithms is transferred to the user. **The assistance provided by these solutions should be swift, accurate, and easily understandable** for P-VI/blindness.

Various studies showed that people with blindness would like to be able to interact with the assistive solutions through voice commands [108] because touch screens are not very efficient for them, and they cannot interact with traditional touch user interfaces as easily as a sighted

person [151]. However, a voice user interface (VUI) has its limitations. For instance, people with blindness prefer to retain the ability to hear the sounds in their surroundings, as this is one of the main sources of perceiving happenings around them. For this reason, it is important to **consider various methods for interaction**. For example, a VUI for the user who is blind could be complemented with haptic feedback to communicate with the user [144]. Due to the aforementioned challenges, it can be beneficial to provide a touch-based interface despite all the challenges that people with blindness have with the dynamics and orientations of such interfaces [152]. This is because touch user interfaces provide more privacy to the user while retaining less of the user's hearing ability.

As we have mentioned earlier, P-VI/blindness can have different needs according to the variation of their disability. For instance, some individuals with VI may prefer enlarging the text size instead of using VUI or screen readers.

Personalization is one of the aspects of assistive solutions that were discussed in [108]. Allowing users to customize according to their level of disability, altering the voice output, and controlling the volume and rate of information delivery, measurement units, warning formats, and text input methods, as well as the ability to adjust the system's volume at any time, were among the customizations users with blindness requested [108].

Besides the software solution, the hardware is also important for the user experience. For instance, deploying the solution as an app on the mobile phone can be challenging if the user is using a white cane/guide dog. In such a situation, both hands of the user will be occupied, which might render the solution uncomfortable for the user.

Moreover, in a system where the sensor device, such as a camera or infrared sensor, must be worn on the body, preliminary findings indicate that a degree of **adjustability of the location of wear** is considered advantageous [153]. For instance, the placement of a sensing device on top of summer clothing may differ from that of a winter coat. There is a reluctance to wear any sensing device on the head, even though the head as a sensor location may provide good results due to its height [153].

Furthermore, a significant advantage of some smart assistive solutions is their **portability**, which allows users to always rely on them. Unlike non-portable solutions such as fixed tactile paving or stationary braille signage, portable assistive devices do not rely on the environment being appropriately accessible for people with a visual impairment. Portable devices can be

used almost anywhere and provide a constant source of assistance, regardless of the accessibility features of the environment.

In addition, it is important to consider the **aesthetics of the assistive device**. Shinohara and Wobbrock [154] suggest that the aesthetics of assistive technology and social acceptance may contribute to stigmatization of users. In [155] it was found that aesthetics seems to have a great influence on how people see and judge users. After interviewing eight P-VI/blindness participants they found out that wearing modern devices that are subtle and have modern aesthetics (e.g. smart glasses) was better accepted by the participants. Therefore, it is an important factor that influences device adoption or abandonment.

To help designers address this challenge, we propose a series of relevant non-functional requirements (NFRs) according to the state-of-the-art regarding assistive solutions for P-VI/blindness, which includes robustness, portability, latency, affordability, accuracy, power consumption, computational complexity, security, and personalization. They are discussed in detail later in the proposed solution (Chapter 6.2). Additionally, we designed our proposed solutions according to the mentioned non-functional requirements to better consider the user experience.

3.4 Scope of the research

This study is going to be focused mainly on the problem of scene understanding and object location for P-VI/blindness in indoor environments. It was decided to focus on indoor environments because of their unique complexities, where traditional outdoor navigation aids (like GPS or environmental cues) are less effective or unavailable. Besides, indoor environments are generally safer than outdoor settings, making them more suitable for the initial testing and implementation of an early-stage solution. To address the mentioned problems, the challenge of considering user needs and the process of selecting appropriate technology will be explored. Additionally, the research will address how to effectively communicate the information provided by the solution to the intended users to consider the user experience. It is also important to mention that the scope of this research is focused on tackling the mentioned issues using technologies that are cost-efficient and have the potential to be commercially affordable for the intended users.

4 Methodology

To address the mentioned problems in the previous section regarding scene understanding for P-VI/blindness in indoor environments, the Design Science Research Methodology (DSRM) is utilized which allows to thoroughly cover the objectives of the thesis while verifying each step and contributions. The DSRM methodology [156], combines two different but complementary paradigms: design science and behavioral science, within the study of information systems. The first, design science, is a problem-solution paradigm that studies how to improve the capabilities of both individuals and organizations in achieving new goals or increasing process efficiency. On the other hand, the behavioral science paradigm is based on natural science research methods and seeks to explain the behaviors of individuals and organizations through theory development and verification. The combination of both creates a complementary research cycle that provides the theoretical knowledge base necessary for research and the creation of artifacts and evaluations that generate knowledge. This methodology was chosen because it aligns with the thesis objectives and contributions: Applying DSRM would facilitate a comprehensive approach to developing effective assistive solutions by combining rigorous academic research with practical, user-centered design principles.

The implementation of these two paradigms converges in the existence of three main elements (represented by boxes in Figure 8), connected by three cyclical elements that overlap them to ensure integration. These main elements are: Environment, Design Science Research, and Knowledge Base, which provide solid foundations for initiating research in the field of information systems. Each element helps to frame the research and ensures rigor in the process while seeking suitable solutions for its environment, as each belongs to a phase of the design within information systems:

1. **Environment Element:** Defines the problem where we will frame the phenomena to be studied within the specific domain of interest. This part defines the People, their roles, and characteristics as users of our system. It also defines the Organization, meaning who or what is behind our design and with what intention, and finally, the Technical System, indicating what technology will be used in the research. In our case it identifies the People, including P-VI/blindness as users of the system, and their roles and characteristics. The Organizational Systems include participants (users of the

solution), stakeholders (such as disability support organizations), and private companies commercializing assistive solutions. Additionally, it outlines the Technical Systems, focusing on portable and wearable devices, and specifies the Physical Environment, which in this case is indoor settings. Finally, it highlights key Problems and Opportunities, such as user-centered design, selecting appropriate technologies, understanding user needs, and developing cost-effective solutions.

2. **Knowledge Base Element:** This element provides the theoretical foundation for the research by incorporating Scientific Theories and Methods, such as state-of-the-art analysis, functional requirements, and non-functional requirements. It also draws from Experience & Expertise, including user research results and an analysis of existing assistive solutions. Furthermore, it includes Artifacts & Meta-Artifacts—design products and processes—such as solutions for visual impairments/blindness, frameworks for assistive solution design, and lessons learned from testing these solutions.
3. **Design Science Research Element:** This element frames the processes of designing and constructing specific solutions for the environment based on insights from the knowledge base. It involves Building Design Artifacts & Processes, such as creating frameworks for analyzing and designing assistive solutions, developing wearable prototypes for research purposes, and producing portable smartphone-based products for commercial use. The process also includes Evaluation, where frameworks are applied to existing solutions and tested with blind or blindfolded users. The iterative nature of this element is emphasized by its integration with the Relevance Cycle, Rigor Cycle, and Design Cycle, ensuring alignment between practical needs, theoretical foundations, and design iterations.

This research methodology proposes three cycles that facilitate feedback among the three main elements: Environment, Design Science Research, and Knowledge Base. These three cycles combine the two paradigms proposed in the methodology and are applied at different stages of the research [34]. These cycles are:

- **Relevance Cycle:** Provides the research with information about the application domain and ensures that the research aligns with it and considers its requirements and constraints.

- **Rigor Cycle:** Supplies the research with a corpus of knowledge, providing information that is incorporated into the design and construction of an artifact or process.
- **Design Cycle:** Develops and evaluates the solution, in the form of an artifact or process, cyclically until the set objective is achieved.

These cycles are interconnected through the design cycle. On one hand, it provides information that contributes to the knowledge base through the rigor cycle, and on the other, the relevance cycles add improvements or advancements to the environment where the design will be applied. Figure 8 shows the research framework diagram defined in [156], contextualized for the research we developed in this work. The following sections will describe in more detail the information flow we have followed in the research.

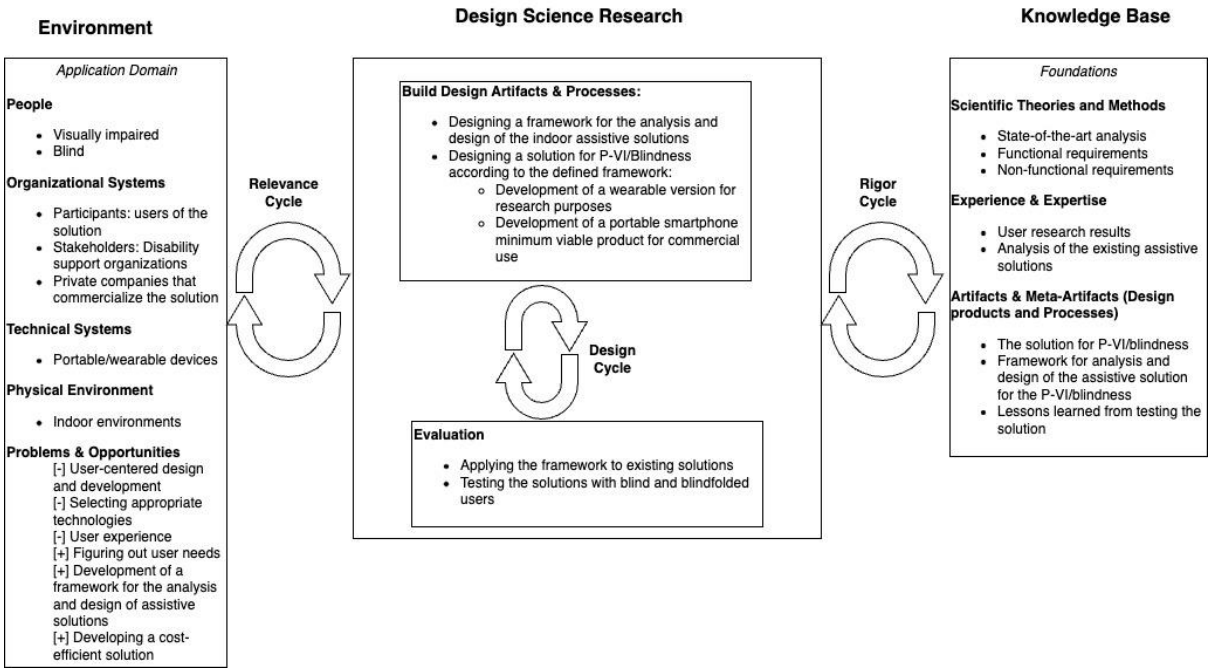

Figure 8 - Design science research methodology adapted to the context of the thesis

4.1 Relevance cycle

The relevance cycle connects the research environment with design science activities, aiming to identify opportunities and problems. Initially, it gathers information about people, physical spaces, and technology, and later, it provides a context for the design and evaluation.

In our Environment, the people are going to be P-VI/blindness, the organizational systems are going to be the users of the solution, disability support organizations, and private companies

that commercialize the solution. The technical systems are going to be based on computer vision, affordable, portable/wearable devices to help in scene understanding, and the physical environment will be any indoor environment, as mentioned in Figure 8. The problems and opportunities define the basis for the design of the research. Each of the problems are addressed in the process of relevance cycle:

- Considering user needs in the development is addressed in the user research.
- Choosing appropriate technologies for the development is addressed in the technology selection chapter.
- The experience of the user when using the solution is addressed in non-functional requirements and the proposed framework section.

4.2 Rigor cycle

This part of the methodology focuses on unifying the scientific knowledge base with the design process, providing the necessary foundations for rigorous research within design science. In this part of the methodology, the state of the art is analyzed and the artifacts and processes found in the domain are defined. Once the developed study generates knowledge, it is transferred back to the knowledge base.

During this cycle, first, an in-depth state-of-the-art analysis has been undertaken. After that a list of use cases were defined that are intended to address the unique challenges associated with understanding indoor scenes. Afterwards, according to the use cases, functional and non-functional user requirements were defined to ensure that the proposed solutions would be effective and practical for the target users. Functional requirements are the capabilities, behavior, and the information that the system requires [108]. On the other hand, non-functional requirements focus on quality constraints like usability, reliability, availability, performance etc. [157]. According to the knowledge gathered, the framework for the analysis and design of assistive solution is designed and added to knowledge base.

Additionally, various technologies (hardware and software) have been evaluated and compared for the development of the assistive solution. The advantages and disadvantages of each approach is addressed to facilitate the design decisions. Finally, the artifacts and meta-artifacts which include the designed solution, the framework and the lessons learned from the testing will be added to the knowledge base.

4.3 Design cycle

The internal design cycle is the core of any design science research project. This cycle involves rapidly iterating between constructing an artifact, evaluating it, and using feedback to further refine the design. This process generates design alternatives and evaluates them against requirements until a satisfactory design is achieved. The requirements come from the relevance cycle, while the design and evaluation theories and methods are sourced from the rigor cycle. Despite its dependence on the other two cycles, the design cycle is where the primary work of design science research occurs. It is essential to understand both its reliance on the relevance and rigor cycles and its relative autonomy during the research execution.

In this cycle, a framework for the design and analysis of the indoor assistive solutions have been developed according to the user requirements. Afterwards, two different solutions are developed according to that architecture considering two different purposes. One for evaluating different technologies and discovering the potentials of different technologies (hardware and software) for the development of a portable assistive solution for the P-VI/blindness. The other one is developed to explore the possibilities of developing a commercial solution for the intended users.

The framework is applied to the existing solutions for evaluation and the solutions are designed based on the framework and then tested with the end users.

5 User research

During the state-of-the-art analysis, we managed to figure out the needs of the P-VI/blindness to some extent. However, to further explore the requirements, and possibly undiscovered user needs, for developing a solution specifically for indoor scene understanding, we carried out a series of semi-structured interviews with people that have different levels of visual impairment and blindness. 8 participants were selected living in different parts of the world with different levels of blindness. Interviews were aimed to investigate the ways in which P-VI/blindness people currently undertake their daily tasks and overcome their challenges, and discover the needs of the blind people in indoor environments regarding scene understanding. The eight participants volunteered after contacting more than 50 blind users through X (previously Twitter) platform due to the ease of using social media platforms to connect with potential participants. This approach allowed us to quickly connect with potential participants who were active online and willing to share their experiences. However, it is important to note that their selection was not guided by a formal sampling strategy or specific quotas for user types, given the difficulty of reaching a large number of P-VI/blindness. Consequently, their diverse backgrounds and cultural contexts were not explicitly considered in the research. While this reduces the generalizability of the findings, it provided valuable initial insights into user needs.

5.1 Methods

According to the state-of-the-art analysis, a list of questions was proposed for the semi-structured interview to explore the area of scene understanding. This included how P-VI/blindness build mental models, perceive different modalities such as sound and vibration for scene description and locating objects in the environment. Moreover, during the interview some scenarios were proposed to figure out how such solutions could assist them in various situations. The list of interview questions is in Appendix 6.

The interviewees were found through social media (mainly Twitter and Instagram). The participants level of visual impairment/blindness were as follows:

- Participant 1 (low vision, night blindness).
- Participant 2 (very low vision).
- Participant 3 (blind - only light perception).
- Participant 4 (congenitally blind).

- Participant 5 (acquired blindness during high school).
- Participant 6 (congenital blindness).
- Participant 7 (congenital blindness).
- Participant 8 (lost vision at 14, right eye fully blind, left eye 20 percent blind. Sensitive to light).

All of the interviews were held online through Zoom or Google Meet from October to December 2023. The sessions were recorded with the consent of the user and later investigated to discover all the important details. The languages of the interviews were Farsi and English since interviewees were from English and Farsi speaking countries such as the UK, USA and Iran.

Initially, participants were asked to introduce themselves and talk about their lives, activities, visual impairment, and the role of scene understanding and navigation in their lives. Topics raised in this introduction were then examined in more detail. Other topics covered, though not all of them were relevant to all participants, were as follows:

1. The use of current assistive solutions.
2. Orientation and mobility in different indoor environments.
3. Different output modalities of the assistive solutions.
4. Most difficult tasks they undertake in their daily tasks.
5. Frequency and the kind of objects they lose.
6. Recommendations systems that could inform them about hazards or interesting objects in the environment.

Once the information was obtained from the interviews, a thematic analysis was applied to extract the common points that were mentioned by various users.

It is important to mention that the user research protocol was approved by the Ethics Committee of Universidad Politécnica de Madrid under reference number 2022–077.

5.2 Results

The findings are organized into sections based on key themes identified during the initial thematic coding of the results. The full table of user research results is in Appendix 7.

5.2.1 Existing assistive solutions

For scene understanding, almost all of the participants were using the help of another person by asking, or moving around the place carefully using a white cane or a guide dog. Besides the conventional solutions, all users used digital assistive solutions such as AIRA [104], Google Maps, Seeing AI [96], Be My Eyes [105], etc. for navigation, locating objects and other tasks such as color detection and text reading. However, although these solutions can increase the user's independence, they still have some limitations that were highlighted by the interviewees. As P#2 expressed her opinion about Seeing AI: *"It has a long long way to go. It is not good for detecting a lot of particulars in the scene. I don't use scene understanding that frequently."* Or P#7: *"I use iOS view finder but I think still it doesn't have enough accuracy. It mistakenly detects a shirt as a towel! Some apps provide color detection but I don't use them because it makes a lot of mistakes. Additionally, scene descriptions do not describe what we want! For example Google Lookout. Seeing AI is better though."* Another participant expressed discomfort in using solutions such as AIRA or Be my eyes because an actual person assists through your phone's camera. P#5 *"Aira felt uncomfortable because an actual person was in my phone knowing my exact location. I know it is supposed to be trustworthy and stuff but I prefer not to use actual person services it's just my anxiety over that. I prefer to ask someone around or just use internet.".* The answers of the users are shown in Table 5.

	P#1	***P#2***	***P#3***	***P#4***	***P#5***	***P#6***	***P#7***	***P#8***
Navigation	white cane	white cane	white cane	white cane	Mostly guide dog	white cane	white cane	Wheelchair user
Scene understanding	Ask others, move around carefully	Ask others, move carefully with white cane	Ask others, move around carefully	Ask others, move around carefully	Ask others, move around carefully	Ask others, move around carefully	Ask others, move around carefully	Ask others for help
Object finding	Ask others, move around carefully	Ask others, Aira	Ask others, move around carefully	Ask others, move around carefully	Using guide dog, move around carefully, ask others	Ask others, move around carefully	Ask others, move around carefully	Ask others for help
Digital services	Google maps,	GPS apps, Braille	Seeing ai (barcode reader and	Nearby Explorer,	Seeing AI, Google	dot walker, Google	Seeing AI, lookout, VoiceOve	iOS VoiceOver

	screen reader	display, voice over, be my eyes, Aira, Seeing AI	OCR), BeMyEyes, Google maps walking mode, blind square	Google maps walk mode	maps, Compass app, Aira	maps, Envision AI	r, image description	

Table 5 - User research (existing solutions)

5.2.2 Modality

Users were asked about their preferred feedback modality for assistive solutions including sound, vibration and voice output. They were also asked what would be the most convenient way of informing them about the distance and location of objects. They mentioned their preferred units of measurement for object position (angles, cardinal directions and clock face) and object distance (meter, feet, steps, frequent beeping/vibration). The answers regarding the distance unit were variant among participants. However, almost all of them preferred either frequent beeping (which gets more frequent when the user gets close to the object), vibration or both for understanding the distance of an object from them. Also, a couple of them noted that it is very important not to block the hearing of the P-VI/blindness because they perceive a lot of information through their hearing sense. P#7 said that *"Using headphones is very tricky for us because headphones block our hearing. Only with bone-conduction headphones I feel comfortable."* P#1 also expressed that he would like to use bone conduction headphones that do not block his hearing. The answers are shown in Table 6.

	P#1	***P#2***	***P#3***	***P#4***	***P#5***	***P#6***	***P#7***	***P#8***
Beeps	✓	✓	X	✓	✓	✓	✓	✓
Vibration	X (afraid of missing)	✓	X	✓	✓	✓ But preferred beep for more accuracy	✓	✓
Distance unit	Meter, Angle	Feet, Cardinal direction	feet, inch, clock method, using reference object	meters, clock face method, reference objects	Steps	Meters	Meters, clock face method	preferred beeping over meters

Augmented audio	unfamiliar	X (prefers verbal directions)	X	✓ (I have played augmented audio games and I think it can be useful.)	X (never used)	X	X	X

Table 6 - User research (feedback modality)

5.2.3 Scene understanding

Participants were also asked about the information they need to perceive for scene understanding. Knowing about objects, color detection, text reading, obstacle detection and scene type were among the most needed information. Participants expressed which kind of obstacles are more challenging for them to avoid and how their current tools are not able to detect those obstacles. For instance, P#1 said that he would need obstacle detection specially for moving obstacles. P#2 mentioned that guide dog and white cane are not useful for the detection of upper body obstacles. Moreover, they expressed how they prefer to receive information about the existing objects in the scene. P#7,2,4 expressed that they would like to receive information on demand and in a hierarchical format to avoid cognitive overload. For example, P#2 noted *"First having general information and being able to dig into it would be good, like knowing what is on it[table] or the size of it, color of it, and so on, having a layered approach."*. Table 7 shows the information that the participants needed to receive from their surrounding environment.

	P#1	***P#2***	***P#3***	***P#4***	***P#5***	***P#6***	***P#7***	***P#8***
Objects	✓moving objects and stuff on the floor	✓ Hierarchical information	✓ Stairs	✓ (Hierarchical information, windows, kitchen objects)	✓	✓	✓ Hierarchical information	✓
Text reading	✓	✓	✓	✓	✓	✓	✓	✓
Obstacle detection	✓(specially moving obstacles)	✓ (upper body obstacles)	✓ (obstacles like columns)	✓	✓	✓	NA	✓
Scene type	N/A	✓	✓	✓	✓	✓	✓	NA

Table 7 - User research (scene understanding)

5.2.4 Recommendation system

It was also explored if users would like to have a recommendation system in the solution that informs them about hazardous situations or interesting objects (e.g. TV, Radio, vending machines, empty seats) in a new environment. P#4 commented that *"It can be very good for finding empty seats in metro stations".* However, the idea of informing them about hazards was more interesting for the participants. Three of the interviewees, P#5, P#3 and P#1, did not support the idea of having a recommendation system for the interesting objects and two of them did not provide any answers about that. Figure 9 shows the number of participants who agreed on including hazards and interesting objects in the recommendation system.

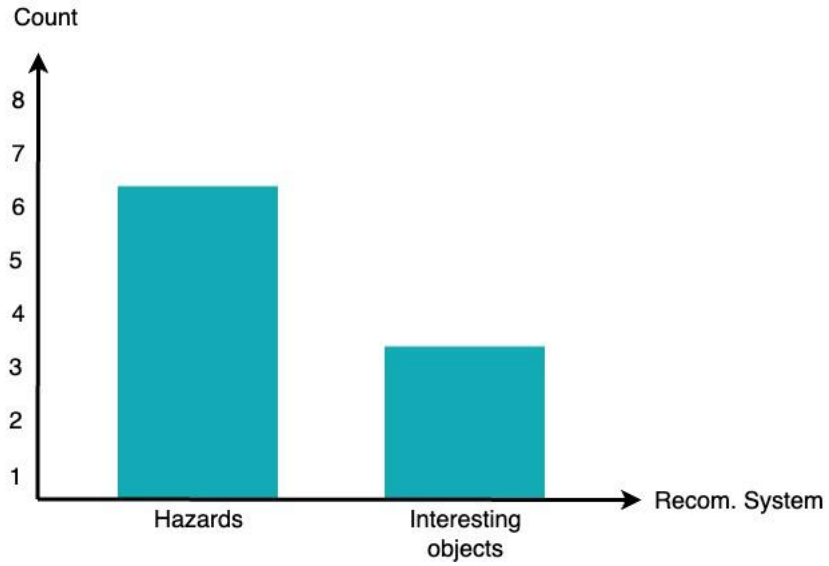

Figure 9 - User research (recommendation system)

5.2.5 Losing objects

Participants were asked how frequently they missed items in their daily lives. Those who lived with sighted housemates more frequently misplaced objects in comparison with the ones who lived alone or with another person with a VI/blindness. Four out of eight participants noted that they lose objects every day. P#6 commented *"I lose objects every day, I live in a dormitory and in my own home with my family. At home, especially in the kitchen, my mom always changes the place of things. I love cooking, but I don't use my mom's kitchen because of that".* More information regarding the type of lost objects is in Table 8.

Objects	Count	Participant
bank cards	2	P4,P6
body spray	2	P7,P6
books	1	P2
charger	2	P5,P4
cigarettes	1	P4
comb	1	P6
dog toys	1	P5
flash memory	1	P4
fork	1	P4
glasses	1	P1
glasses case	1	P1
hanger	1	P7
hat	1	P6
headphones	3	P5,P4,P2
keyboard	1	P2
keys	4	P5,P4,P6,P7
lighter	1	P4
nail cutter	1	P7
paper tissue	1	P6
phone	3	P1,P4,P5
pills/medicine	1	P4
remote controller	2	P7,P8
shoes	1	P5
spoon	1	P4
wallet	2	P5,P4
watch	1	P4
water bottle	1	P5

Table 8 - Most frequent lost objects

5.2.6 Most complex tasks

Regarding the most complex tasks, finding objects and directions were the most complicated challenges that the participants faced in their daily lives. These tasks included finding objects that others move (P#1), dealing with images and audios with no descriptions (P#2), finding stairs (P#3), finding registers in big stores (P#3), cooking (P#4), finding stuff in the bathrooms or kitchens (P#5), reading mails (p#8) etc. Results are detailed in Table 9.

	P#1	***P#2***	***P#3***	***P#4***	***P#5***	***P#6***	***P#7***	***P#8***
Most complex tasks	finding objects when others move them	dealing with images and audios with no description	finding stairs, finding registers in big stores	cooking, finding directions indoors	Finding stuff in the new bathrooms/kitchens, cooking, using appliances, little steps in American houses	cooking, finding stuff in the new bathrooms/kitchen	moving in the streets	reading paper mails

Table 9 - User research (Complex tasks)

5.3 Conclusion

Overall, the user research provided valuable insights into the needs and challenges faced by P-VI/blindness in indoor environments. Participants currently rely on a mix of traditional tools (e.g., white canes, guide dogs) and digital assistive technologies (e.g., Seeing AI, Be My Eyes, Google Maps). However, these solutions often lack accuracy, fail to address specific scene understanding needs, and sometimes create discomfort due to privacy concerns.

Key findings revealed that participants require better obstacle detection (including moving and upper-body obstacles), hierarchical information delivery to avoid cognitive overload, and non-intrusive feedback modalities such as sound or vibration. Bone-conduction headphones were particularly favored for maintaining auditory awareness. Participants also expressed a need for solutions that provide on-demand, layered information about objects, text reading, and scene classification.

Frequent challenges included locating misplaced objects, navigating unfamiliar spaces, and performing complex tasks like cooking or finding specific items in shared environments. While participants supported the idea of hazard alerts in recommendation systems, interest in identifying “interesting objects” was mixed.

These findings emphasize the need for more accurate, user-centered assistive technologies that address gaps in scene understanding while ensuring usability and comfort for P-VI individuals.

6 Proposed framework

To address the challenges mentioned in the problem statement chapter, such as proper integration of user needs in solutions, selection of appropriate technologies, and providing proper feedback to the user, a framework was designed that consists of:

- The various use cases of an assistive solution for scene understanding in indoor environments.
- The list of functional and non-functional requirements to be fulfilled by an assistive solution for scene understanding in indoor environments.
- A general reference architecture for assistive solutions for scene understanding in indoor environments.
- A guideline for selecting appropriate technologies for the design and development of such solutions.

User-centered design and development: In response to the lack of a user-centered approach in current systems, we contribute to a deeper understanding and characterization of the context of use—covering users, tasks, and the environment. This is achieved through the analysis of the possible multiple use cases for assistive solutions in indoor scene understanding.

Moreover, recognizing the absence of comprehensive design guidelines that link user needs to system requirements and architecture, we present a comprehensive set of functional and non-functional requirements, a reference architecture, and a mapping between requirements and architectural components.

Selecting appropriate technologies: To address the lack of clear guidelines for the selection of appropriate technologies, we contribute a detailed technology analysis. This analysis is applied to facilitate the technology selection process for developers.

All the elements in the framework will be applied to our proposed solutions

6.1 Use cases

This section presents a set of structured use cases that help to visualize and categorize the diverse needs and real-world challenges faced by the P-VI/blindness in indoor settings. They have been elaborated by collecting and systematizing the results of the user interviews and our

previous literature analysis [13]. The identified use cases are divided into four main categories: “Scene understanding”, “Object location”, “Obstacle avoidance” and “Text reading”. The “Scene understanding” use cases are focused on situations in P-VI/blindness gets to an unfamiliar indoor environment and would like to obtain information about it and the objects present in it. “Object location” is related to situations in which the user is looking for specific objects (e.g., book, clothes, seats, etc.). In “Obstacle avoidance” use cases, the user wants to avoid obstacles on a trajectory. Lastly, “Text reading” use cases are for situations where the user wants to detect and understand a text in the environment.

6.1.1 Scene understanding use cases

6.1.1.1 Wide indoor scene (room)

In the following use cases, it is assumed that the user has entered an unfamiliar living room of an apartment and would like to have a better understanding of the surroundings and the existing objects/living entities with the help of the assistive solution. Each use case defines a possible scenario that might occur while the user is in the indoor environment. In the following use cases, the objects' distance is considered to be at least 1 meter from the user. This is because bigger objects need to be far enough from the camera to be completely included in the image obtained from the scene.

Use case 1: Scene identification

> As a user, I want to know what type of room I am in when I enter the room, so that I can be sure I entered the correct room.

Use case 2: Scene description (static objects and standing user)

> As a user, I want to get informed about the static (not moving) objects (e.g. chairs, TV, books etc.) present in the environment while I am standing still, so that I can understand what and where objects are in the room.

Use case 3: Scene dynamics description (moving objects and standing user)

> As a user, I want to get informed about moving objects (e.g. humans or pets) present in the living room so that I can be more careful about them if I move.

Use case 4: Static scene description for navigation (static objects and moving user)

As a user, I want to get informed about the objects present in the current environment, while I am moving, so that I can form a better mental map of the environment.

Use case 5: Dynamic scene description for navigation (moving objects and moving user)

As a user, I want to get informed about the objects present in the current environment that are moving (e.g. humans or pets), while I am also moving, so that I can form a better mental map of the environment and at the same time be more careful about moving objects.

Use case 6: Scene object type description (static/moving objects and standing/moving user)

As a user, I want to be informed about objects of the same type (e.g. three books) in the environment so that I can distinguish them and know how many objects of the same type exist in the environment (e.g. five chairs).

6.1.1.2 Small indoor scene (area within hands reach)

In the following use cases, it is assumed that the user is sitting at a table and would like to know about the existing objects on the table with the help of the assistive solution. Each use case defines a possible scenario that might occur in this setting. The objects in these use cases are less than 1 meter away from the user (within hands reach).

Use case 7: Scan items on a table (static objects and sitting user)

As a user, I want to get informed about all objects (food, drinks, plates, fork etc.) on the table so that I can decide if there is any object of interest.

Use case 8: Scan specific object type on a table (static objects and sitting user)

As a user, I want to know if there is any instance of a specific type of object (forks, for instance) on the table.

6.1.2 Object location use cases

6.1.2.1 Wide indoor scene (room)

In the following use cases, it is assumed that the user has entered an unfamiliar living room of an apartment and wants to locate a specific object with the help of the assistive solution. The objects in the scene could be large (i.e., easily visible in the image) or small (i.e., occupying fewer pixels in the image). We will discriminate between objects that are close (less than one

meter away) and far (more than one meter away). In the wide indoor scene category only far objects will be considered, while close objects will be included in the small indoor scene category. However, the distance threshold could be adjusted based on the user's requirements. These metrics are intended to provide a more tangible understanding of the use cases.

Each use case defines a possible scenario that might occur while the user is undertaking the task.

Use case 9: Finding a big far static object (standing user)

As a user, I want to know the location of the sofa in the living room so that I can go there and seat on it.

Use case 10: Finding a big far moving object (standing user)

As a user, I want to know the location of a person moving in the living room.

Use case 11: Finding a small far static object (standing user)

As a user, I want to know the location of a bottle of water (>1 meter away) while I am standing in the living room so that I can drink water from it.

Use case 12: Finding a big far static object (moving user)

As a user, I want to know the location of the sofa while moving in the living room so that I can go there and seat on it.

Use case 13: Finding a small far static object (moving user)

As a user, I want to know the location of a remote controller while I am moving around the living room so that I can turn on the air conditioner.

Use case 14: Finding a small far moving object (moving user)

As a user, I want to know the location of a kitten while I am moving around the living room so that I can orient myself to it and call it.

6.1.2.2 Small indoor scene (area within hands reach)

In the following use cases, it is assumed that the user is sitting at a table and would like to locate existing objects on the table with the help of the assistive solution. It is also assumed that all objects are small and static, as big or moving objects are not usually found on a table.

The distance between the user and the object in the scene is less than one meter, so that the user could reach the desired objects without walking, just by extending their arm. Each use case defines a possible scenario that might occur.

Use case 15: Finding a small static object (sitting user and static objects on a table)

As a user, I want to know the location of a salad bowl while I am sitting at the table so that I can get some salad.

Use case 16: Finding a small static object with a specific color (sitting user and static objects on a table)

As a user, I want to know the location of a blue book while I am sitting at the table so that I can read it.

Use case 17: Finding the color of any desired small object (sitting user and static objects on a table)

As a user, I want to know the color of a crayon I picked from the table.

6.1.3 Obstacle avoidance use cases

In the following use cases, it is assumed that the user is moving around a living room and wants to avoid colliding with obstacles. Each use case defines a possible scenario that might occur while the user is undertaking the task.

Use case 18: Obstacle avoidance (small static obstacle)

As a user, I want to be informed about small static obstacles (e.g. small vase) on my way while I am walking around the living room so that I can move around safely.

Use case 19: Obstacle avoidance (small moving obstacle)

As a user, I want to be informed about small moving obstacles (e.g., a cat) on my way while I am walking around the living room so that I can move around safely.

Use case 20: Obstacle avoidance (big static obstacle)

As a user, I want to be informed about big static obstacles (e.g., sofa) on my way while I am walking around the living room so that I can move around safely.

Use case 21: Obstacle avoidance (big moving obstacle)

As a user, I want to be informed about big moving obstacles (e.g., person) on my way while I am walking around the living room so that I can move around safely.

Use case 22: Hazard warning

As a user, I want to be informed about potential hazards around me (e.g., staircase, slippery floor) on my way while I am walking around the living room so that I can move around safely.

6.1.4 Text reading use cases

Use case 23: text detection

As a user, I want to be able to know if there is any text written on a given surface.

Use case 24: text reading

As a user, I want to be able to know the content written in a magazine so that I can understand it.

It is important to note that the differences between moving and standing users, as well as big and small objects, can have significant impacts on the performance of the system for scene understanding. In situations where the user is standing still or sitting (e.g., Use case 2, 3, 7, 8, 9, 10, 13 and 14), the system can focus on detecting and analyzing the static objects in the environment. The user's stationary position allows for more accurate and reliable object detection as there are fewer variables to consider.

However, when the user is moving (e.g., Use case 4, 5, 11, 12, 18, 19, 20, 21 and 22), the system needs to account for the user's motion while detecting and tracking objects. This introduces additional complexity as the system must handle object detection in dynamic scenes. It needs to continuously track the user's position, adjust object detection algorithms accordingly, and provide updates about the objects' locations. The system’s performance may be affected by factors such as motion blur, occlusions, and changes in lighting conditions caused by the user's movement. However, a static user might be moving the camera in order to scan the environment, so these problems could also affect to some extent the static use cases.

Moreover, object size is also important. When dealing with large objects (as in Use case 9, 10, 12, 20 and 21), the algorithm must accurately locate and track these objects. Larger objects are more likely to stand out in a scene.

Small objects, on the other hand (e.g., Use case 11, 13, 14, 15, 16 and 17), require the algorithm to detect and localize them accurately. Small objects usually occupy fewer pixels in an image. This results in limited visual details available for object detection models to identify and differentiate these objects. Small variations or distortions in these few pixels can drastically affect the detection accuracy [158]. The algorithm should be capable of capturing precise details and adjusting its detection and localization methods to accommodate small-scale objects within reach of the user. It is also important to mention that as the distance of the object from the camera affects the object's size in the image, a big object far from the user could occupy just a few pixels of the image and vice versa.

Moreover, small objects may often require the user to perform finer and more precise movements to reach or avoid them. The system needs to provide accurate real-time feedback to allow the user to adjust their movements accordingly. Therefore, not only detection but the feedback mechanism for navigation also needs to be highly precise and reliable for small objects. This is because smaller objects may be covered or surrounded by other objects, making it difficult to access them. For example, the user wants to pick a fork, but there is a cup nearby that the user's hand may collide with. In [149], [159] discussed the importance of the feedback mechanism and the response time of assistive solutions in the case of object detection and obstacle detection. They also discussed various modalities that can be used such as vibrations, audible alerts, or echo waves. Nonetheless, according to our user research, we suggest that different users should be allowed to adapt the feedback mechanisms for navigation according to their personal preferences and each use case. The full list of all scenarios is in Table 10.

	User need	**Location**	**User position**	**Object/s**	**Object size**	**Distance**
Use case 1	Scene recognition	Room	Standing	Static	Small & Big	>1 meter
Use case 2	Scene description	Room	Standing	Static	Small & Big	>1 meter
Use case 3	Scene description	Room	Standing	Dynamic	Small & Big	>1 meter
Use case 4	Scene description	Room	Moving	Static	Small & Big	>1 meter

Use case 5	Scene description	Room	Moving	Dynamic	Small & Big	>1 meter
Use case 6	Scene description	Room	Moving/Standing	Static/Dynamic duplicated objects	Small & Big	>1 meter
Use case 7	Scene description	Table	Sitting	Static	Small	<1 meter
Use case 8	Scene description	Table	Sitting	Static specific object	Small	<1 meter
Use case 9	Object location	Room	Standing	Static	Big	>1 meter
Use case 10	Object location	Room	Moving	Static	Big	>1 meter
Use case 11	Object location	Room	Moving	Static	Small	>1 meter
Use case 12	Object location	Room	Moving	Static	Big	>1 meter
Use case 13	Object location	Room	Standing	Static	Small	>1 meter
Use case 14	Object location	Room	Standing	Moving	Small	>1 meter
Use case 15	Object location	Table	Sitting	Static	Small	<1 meter
Use case 16	Color detection	Table	Sitting	Static	Small	<1 meter
Use case 17	Color detection	Table	Sitting	Static (any object)	Small	<1 meter
Use case 18	Obstacle avoidance	Room	Moving	Static	Small	<1 meter
Use case 19	Obstacle avoidance	Room	Moving	Dynamic	Small	<1 meter
Use case 20	Obstacle avoidance	Room	Moving	Static	Big	<1 meter

Use case 21	Obstacle avoidance	Room	Moving	Dynamic	Big	<1 meter
Use case 22	Hazard warning	Room	Moving	Static	Small/big	
Use case 23	Text detection	Room	Standing	Static	NA	<1 meter
Use case 24	Text reading	Room	Standing	Static	NA	<1 meter

Table 10 - Use cases

6.2 User requirements

In the next two sub sections, we present the user requirements for assistive solutions that have been obtained from the review of previous research. They have been abstracted from the extensive list of requirements extracted from literature (Appendix 3). The requirements are divided into functional and non-functional. Functional requirements are the capabilities, behavior, and the information that the system requires [108]. On the other hand, non-functional requirements focus on quality constraints like usability, reliability, availability, performance etc. [157].

6.2.1 Functional requirements

The functional requirements have been linked to the use cases defined previously in the use case section.

FR 1: The system must have a command interpreter that allows the user to choose between the options provided in the system. (Use case 1-24)

FR 2: The system must obtain scene data to allow the user to scan the environment/text through a camera. (Use case 1-24)

FR 3: The system must recognize the scene type. (Use case 1)

FR 4: The system must describe the objects that exist in the scene. (Use case 2-7)

FR 5: The system must detect the desired object and locate its position in the environment. (Use case 8-15)

FR 6: The system must detect the color of a desired object in the environment. (Use case 16-17)

FR 7: The system must detect the obstacles/hazards that are on the user's path. (Use case 18-22)

FR 8: The system must let the user scan text and know the written words. (Use case 23-24)

FR 9: The system must continuously track object locations as the user moves through the environment. (Use case 4, 5, 10, 12, 13 and 14)

FR 10: The Text detector module must detect the presence of text in the scene. (Use case 23-24)

FR 11: The system must provide feedback based on user preferences (volume, speed, units, etc.) via audio and/or haptics based on user needs. (Use case 1-24)

6.2.2 Non-functional requirements

Non-functional requirements are just as important as the functional requirements since they describe the required quality of the operations handled by the system.

According to the discussion in [108], [144], [160] with several associations of persons with visual impairment and our own interviews, the following non-functional requirements should be addressed in a portable assistive device for the people who are blind.

NFR 1: Robustness - the system should not be impacted by the scene dynamics or illumination. It should be dependable under various circumstances, including water, knocks, and bumps, and need little maintenance. This extends to the capability of functioning seamlessly in diverse environments, including challenging scenarios such as underground locations with no internet connectivity.

NFR 2: Portability - the device should be light, comfortable, and ergonomic.

NFR 3: Latency - The system's latency should be fast enough according to the use case. For instance, when the user requires a scene description while standing still, a slight delay of a few seconds in providing feedback may be acceptable, allowing the system to thoroughly analyze the environment and provide accurate information. However, in the case of obstacle avoidance, the system should operate with millisecond-level responsiveness. This is crucial for ensuring

the user's safety, especially since sensors like LiDAR or ultrasonic sensors commonly used for this purpose can deliver feedback in under a second.

NFR 4: Affordability - the price of the device should be reasonable. (More information about the price of acquisition is provided in section 3.7)

NFR 5: Accuracy - The system must be able to accurately recognize and warn about objects/texts and obstacles. However, the level of accuracy is scenario dependent. For example, in cases where obstacle avoidance is required, the system's accuracy must be very high. This means that it should be capable of reliably detecting and alerting about obstacles in its surroundings to ensure the safety of users. On the other hand, when it comes to object detection, the system's accuracy requirements are slightly different. While it is ideal for the system to detect and recognize all objects present in a given scene, it is understood that achieving 100% detection may not always be feasible or necessary.

NFR 6: Power consumption - The hardware resources that rely on a battery supply should be used by the system in an effective way.

NFR 7: Computational complexity - The system should not impose a heavy computational complexity load on the portable hardware. In other words, the system should be designed and optimized in a way that it doesn't overly burden or strain the processing capabilities of the hardware it runs on. This is especially important for portable devices with limited computing power, such as smartphones or tablets, where efficient resource usage is critical to ensuring smooth and responsive operation while avoiding excessive battery drain.

NFR 8: Security – It is very important that the system guarantees the privacy of the users with blindness and the bystanders who are exposed to the system. In [142] was found out that both users with blindness and bystanders have significant concerns about the information provided by the system. For example, bystanders may feel uneasy if assistive technology extends a sighted person's field of view. On the other hand, the data captured by the cameras on the device must be encrypted so that the user with blindness feels comfortable using the device especially in private environments.

When incorporating text-to-speech and speech-to-text modalities, the privacy concerns are further amplified. The conversion of text to speech and vice versa involves processing sensitive information, and potential risks of information disclosure.

NFR 9: Personalization - The assistive solution must be adaptable to the user's specific needs and preferences in different use cases. This means that the user should be able to adjust various settings (e.g. feedback volume, distance thresholds and modalities) of the solution to optimize its performance for their individual requirements [161].

6.2.3 Reference System Architecture

According to the functional and non-functional requirements, a high-level reference architecture for scene understanding assistive solutions is proposed which is presented in Figure 10. The architecture is organized in a way that facilitates comprehension and implementation of such systems. It serves as a comprehensive reference, providing a high-level overview that aids in understanding the system's major components and their interactions. We will also describe how these components meet the functional and non-functional requirements.

The architecture consists of several key modules, each serving a specific purpose in the scene understanding process. In our proposed architecture, the system enables user interaction through a command interpreter (e.g. voice user interface (VUI), touch screen, keyboard) (FR 1), allowing users to issue commands that the system processes to provide relevant responses. Additionally, The Scene data obtainer (e.g. camera, sensor) provides the input data from environment to the system (FR 2).

The Object locator module, comprising an Object detector model and a Distance predictor, is responsible for identifying the position and type of objects within the environment (FR 5). Additionally, the Scene recognizer module works in tandem with the Object detection model to determine the scene type based on the detected objects within it (FR 3). This interaction between the modules enhances the system's ability to recognize and understand complex scenes more effectively.

Furthermore, the Scene analyzer and descriptor module provides in-depth information about the objects/colors present in the scene, as well as their semantic relationships (FR 4 & 6). To address the recognition of text, the Optical Character Recognition (OCR) module, or Text detector, is employed. This component can accurately detect and interpret text, enabling the system to convey textual information to the user (FR 8 & 10).

Lastly, the Obstacle detector module plays a critical role in ensuring user safety by identifying obstacles present around the user (FR 7 & 9). This module uses the Distance predictor to determine the proximity of obstacles and provide timely alerts or information, assisting the user

in navigating their environment safely (FR 11). The information generated by the various modules is converted into speech format or beeping sounds/vibration in case of obstacles (so

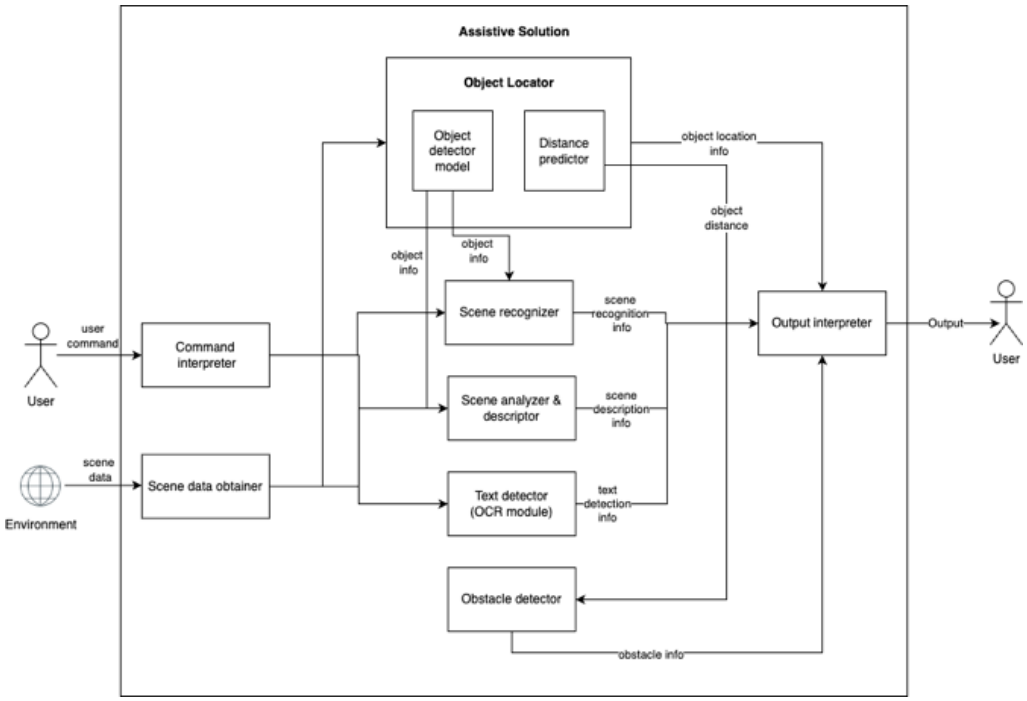

Figure 10 – Solution Architecture: Different modules of the solutions and how they interact with each other

that the beeping/vibration frequency increases/decreases in response to the user's proximity to an obstacle/object.), allowing the system to provide feedback to the users using the Output interpreter module (FR 11). It is also important to note that the quality of each module is highly dependent on the non-functional requirements described. For instance, the accuracy (NFR 5) and latency (NFR 3) of the object detection algorithms used in the modules can highly affect the performance of the system. The impact of the non-functional requirements is discussed in more detail in the next sections.

6.3 Technology selection guidelines

The process of choosing the right technologies for the development of a solution could be a complicated task. During the state-of-the-art analysis we came across many different technologies both for the hardware and software that were used to implement a solution. However, the justification for choosing a specific method for scene understanding (e.g. object detection methods, distance detection sensors/cameras etc.) remained to an extent vague in many solutions. Additionally, it can be confusing for the designers and developers to choose an approach among many different software and hardware options.

To provide a guide for the technology selection process in the design cycle, we compared various approaches for implementing assistive solutions considering various aspects: effectiveness, computational resources, limitations, and acquisition cost of each approach.

Complete information for the different approaches according to our guideline is in Table 11. For each of these approaches we refer to solutions that were implemented using the algorithms and devices mentioned. However, for a few of the approaches we could not find some information such as the computational resources, or reported effectiveness of the method in the corresponding existing solution. In the following sections, each aspect in our guideline is presented.

By assessing the existing solutions using the mentioned aspects, we expect to help designers of new systems to be implemented for specific use cases. The most appropriate approach should be selected considering the priorities and constraints applicable with respect to the different aspects.

6.3.1 Effectiveness

To analyze effectiveness, we compared the **accuracy** (the ratio of correctly detected objects (true positives)) or **average precision (AP)** (metric in object detection that measures the model's precision across different recall levels) of different algorithms, in combination with the **depth coverage range** (measured in meters) of cameras, to show the advantages and the limitations of each approach.

For instance, object detection methods such as YOLOv7 (56.8% AP) have relatively higher AP in comparison with instance segmentation methods such as Detectron2 R101-FPN (43.7% AP) which requires more computational resources too. Instance segmentation is a better approach for finding the exact position of an object in an image.

For depth estimation, there are different options such as stereo vision/LiDAR/structured light cameras and deep learning methods with monocular cameras. Stereo and LiDAR cameras are more common in assistive solutions and have a considerably good coverage range (up to 4 and 6 meters respectively). Deep learning models are an alternative if a depth camera cannot be utilized in the solution for various reasons such as portability or device limitations (e.g. smartphones).

Regarding the obstacle detection technologies, LiDAR and RGB-D cameras have a much better range in comparison with other sensors. The Realsense LiDAR L515 can detect obstacles up to 6 meters away from the user. The ultrasonic and infrared sensors are also acceptable options for detecting close-range obstacles in cost-efficient solutions.

6.3.2 Acquisition cost

The cost of delivering the solution and the cost of acquisition for the end user can be critical. The equipment required to implement solutions with higher effectiveness will result in a higher price.

For instance, the development of an object location use case, using a jetson nano with Realsense LiDAR L515 can cost around 1000$. This price is only for the hardware, and it is also important to consider the costs related to development and maintenance.

Furthermore, computing resources and purchase costs are interconnected. A technique with low computational and acquisition costs could be implemented on a Raspberry Pi 4 computer. A small object detection model (YOLOv7-tiny) or a cloud service could be utilized with a Raspberry Pi 4 (from 35$) and a Pi camera (30$) which does not necessitate a lot of computational resources. However, the effectiveness of a tiny model is worse than a cloud service provider such as Google Vision API or even the normal version of YoloV7 which has 56.8% average precision (AP) which is higher than the tiny version (35.2% AP).

6.3.3 Computational resources

Computational resources mean the hardware and processing power required to run the solutions efficiently. This includes factors such as processing speed, memory, and GPU capabilities. In Table 11, “computational resources” refers to the specific device on which the solution is implemented, such as a cloud server, a device like Jetson Nano, or a smartphone. Each of these devices has different capabilities that affect the speed, accuracy, and cost of running object detection models. High effectiveness comes at a high computational cost and requires more accurate sensors. In some cases, higher precision for object detection models could be achieved using cloud providers. This also decreases the computational load on the user’s device, but it needs constant connection to the internet and also adds additional costs since these providers are not free to use. Our analysis includes technological options for both cost-effective and high performance (costly) approaches.

It's also important to note that low-cost solutions may not necessarily mean inferior quality. Innovations in technology have allowed for the creation of more cost-effective solutions over time. An ideal balance between cost and effectiveness should be sought to ensure that the users get the maximum benefit. It is preferential to develop something that a broad range of users can buy on a limited budget. For example, using YOLOv7 for the required use cases seems like a promising option when it comes to the balance between cost and performance. On the other hand, for the approaches that require more potent and costly hardware, Jetson Nano Developer Kit would be a suitable option.

6.3.4 Limitations

As we discussed in the previous sections, each of the approaches have their limitations, such as the lack of precision or the computational resources they need for running. It is also important to note that the generalizability of solutions might vary in different situations. The methods analyzed in Table 11, primarily function and have the mentioned performance in restricted situations. This is mainly due to the limitations of current scene interpretation technologies. Existing cameras have constraints such as field of vision, size, performance in various lighting conditions, etc. that might affect the performance of the overall approach. In addition, the performance of the software used for scene interpretation, such as various deep learning algorithms for spotting objects, may not be satisfactory in unfamiliar situations because they are trained on a limited number of object categories. In this dimension we highlight the main limitations to be aware of in each approach.

User need	**Use case**	**Approach**	**Effectiveness (Accuracy and Camera range)**	**Computational resources**	**Limitations**	**Acquisition cost**
Scene recognition	Use case 1	patch feature encoding (Non-negative sparse decomposition model (NNSD)[162])	85.40% accuracy on Indoor-67 dataset [86]	Not available	only limited scenes	Not available
Scene recognition	Use case 1	CNN/Hybrid models (e.g. FOSNet [90])	90.37% accuracy on	Not available	need more computational resources	Not available

			Indoor-67 dataset [86]			
Scene recognition	Use case 1	Lightmodels (e.g. SceneRecog [84])	83.33%[84] accuracy	Android Smartphone	detects less objects	Starting from 200$
Scene description	Use case 2 to 5	Object detection (e.g. YOLOv7)	56.8% average precision (AP)[163]	Jetson Nano/ Raspberry Pi 4 according to [164]	computationally heavy	Jetson Nano From 150$ Raspberry Pi 4 from 35$
Scene description	Use case 2 to 5	Object detection (e.g. YOLOv7-tiny)	35.2% AP [163]	Raspberry Pi 4	detects less objects	Raspberry Pi 4 from 35$ + Pi Camera 30$
Scene description	Use case 6,7	Instance segmentation (e.g. Detectron2 R101-FPN model [165])	43.7% AP on COCO dataset [166]	Jetson Nano according to [167]	computationally heavy	Jetson Nano From 150$ + Pi Camera 30$
Scene description	Use case 6,7	Object detection (e.g. YOLOv7)	56.8% AP[163]	Jetson Nano/ Raspberry Pi 4[164]	Might not detect the exact position of small objects	Jetson Nano From 150$ Raspberry Pi 4 from 35$ + Pi Camera 30$
Scene description	Use case 6,7	MLLM	NA	Smartphone according to [105]	Prone to hallucinations	Smartphone starting from 200$
Scene description	Use case 6,7	Google Vision API	60% AP on Open Images Validation set	Raspberry Pi 4 according to [168]	Needs constant internet connection	Raspberry Pi 4 from 35$ + Pi Camera 30$ + API price
Object location	Use case 8 to 17 and 22	Object detection (YOLOv7-tiny) + stereo camera (e.g. Realsense D455)	35.2% AP[163] for object detection	Raspberry Pi 4[164]	not detecting textureless surfaces	Raspberry Pi 4 from 35$ Realsense D455 is 419$

			0.1 m to 4 m camera range[169]			
Object location	Use case 8 to 17 and 22	Object detection (YOLOv7-tiny) + ToF camera (e.g. Intel RealSense LiDAR L515)	35.2% AP [163] for object detection 0.35 m to 6 m camera range [63]	Raspberry Pi 4 [164]	not energy and cost efficient	Raspberry Pi 4 from 35$
Object location	Use case 8 to 17 and 22	Object detection (YOLOv7-tiny) + monocular depth estimation (e.g. MiDAS [170]v2.1 small)	35.2% AP [163] for object detection 13.43 % mean absolute relative error (MARE) for MiDAS v2.1[171]	Android Smartphone according to [172]	not accurate enough in many cases	Smartphone starting from 200$
Object location	Use case 8 to 17 and 22	Object detection (YOLOv7-tiny) + structured light camera (e.g. Intel RealSense SR305	35.2% AP [163] for object detection 0.5 m camera range	Raspberry Pi 4 [164]	requires more computational power than stereo camera.	Intel RealSense SR305 410$
Object location	Use case 8 to 17 and 22	Instance segmentation (e.g. Detectron2 R101-FPN model [165])+ stereo camera (e.g. Realsense D455)	43.7% AP on COCO dataset 0.1 m to 4 m camera range	Jetson Nano according to [167]	computationally heavy, not detecting textureless surfaces	Jetson Nano From 150$ Realsense D455 419$

Object location	Use case 8 to 17 and 22	Instance segmentation (e.g. Detectron2 R101-FPN model [165]) + ToF camera (e.g. Intel RealSense LiDAR L515)	43.7% AP on COCO dataset [166] for instance segmentation 0.35 m to 6 m camera range [63]	Jetson Nano according to [167]	computationally heavy, not energy and cost efficient	Jetson Nano From 150$ Intel RealSense LiDAR L515 880$
Object location	Use case 8 to 17 and 22	Instance segmentation (e.g. Detectron2 R101-FPN model [165]) + Structured light camera (e.g. Intel RealSense SR305)	43.7% AP on COCO dataset 0.5 m camera range	Jetson Nano	require more power than stereo camera	Jetson Nano From 150$ Intel RealSense SR305 410$
Obstacle avoidance	Use case 18 to 22	LiDAR sensor (e.g. Intel RealSense LiDAR L515)	0.35 m to 6 m camera range [63]	Jetson Nano according to [167] or Rasp berry Pi according to [173]	computationally heavy	Raspberry Pi 4 from 35$ Realsense LiDAR L515 is 880$
Obstacle avoidance	Use case 18 o 22	Ultrasonic sensor (e.g. HC-SR04)	Up to 0.4 m [174]	Rasp berry Pi according to[174]	poor accuracy detecting soft or curved objects	Raspberry Pi 4 from 35$ Ultrasonic sensor 8$
Obstacle avoidance	Use case 18 to 22	Infrared sensor[175]	Around 0.2 m to 1.50 m [76]	Rasp berry Pi according to [76]	short range (up to 1.5 m)	Raspberry Pi 4 from 35$ Infrared sensor 10$
Obstacle avoidance	Use case 18 to 22	RGB-D Camera (Intel Realsense D415[176])	Up to 2 m [176]	Jetson Nano according to [177]	computationally heavy	Jetson Nano From 150$

Text reading	Use case 23, 24	OCR algorithm e.g. Google Tesseract [97]	0.87 accuracy on ICDAR 2019-MLT dataset	Rasp berry Pi according to [178]	hard for the blind user to point at the text, can't detect text on blurry image	Raspberry Pi 4 from 35$ + Pi Camera 30$

Table 11 - Solution approaches comparison

Overall, developers and designers of assistive solutions should use the technology selection guidelines to make informed decisions by balancing criteria such as effectiveness, computational resources, limitations, and acquisition cost. These trade-offs should be carefully considered to ensure that the selected technologies meet user needs effectively and economically while accounting for computational resources and inherent limitations.

In order to choose the most appropriate technologies for the development of an assistive solution with the help of the technology selection table, the following steps should be followed:

1. Define the scope by selecting the applicable use cases:
    1. Define the primary objectives and the related functional and non-functional requirements of the assistive solution.
2. Performance requirements:
    1. Assess the expected effectiveness in terms of accuracy and camera range.
3. Hardware and software considerations:
    1. Determine the computational resources constraints:
        1. High-end (costly): e.g. Jetson Nano + Realsense LiDAR for demanding applications.
        2. Lightweight (cost-efficient): e.g. Raspberry Pi + YOLOv7-tiny for economical setups.
4. Limitation management:
    1. Identify limitations of the potential technologies. E.g. smartphones with lightweight models cause reduced object detection capabilities. On the other hand, resource-heavy CNN/Hybrid models imply higher computational demands.
5. Choose an approach:

1. According to the information collected in previous steps, select the most appropriate approach for the development of the assistive solution with the help of table 11.

6.4 Evaluation with the framework – An example

The proposed collection of use cases, the reference architecture, along with the functional and non-functional requirements, and the technology analysis and selection criteria are meant to guide the design, development and evaluation of assistive solutions for P-VI/blindness. To illustrate how a solution could be evaluated according to the mentioned framework, an existing solution (Seeing AI) has been assessed according to the proposed elements. We chose Seeing AI[96] by Microsoft because it is a popular app used by P-VI/blindness all around the world. It also includes many of the requirements that we discussed.

6.4.1 Functional requirements

In terms of functional requirements, the command interpreter (FR 1) is a screen reader (iOS VoiceOver) and the output (FR 11) of the solution is through voice feedback. It receives the environment image as an input from the phone's camera (FR 2). It has a scene scanning feature that provides minimal explanations about an image taken by the user. It provides a very general description (e.g. "This looks like a kitchen") which meets FR 3. In some cases, it also provides brief information about few objects present in the scene (FR 4). Additionally, it offers a person detector which counts faces, and it is also possible to introduce specific faces by taking some photos of a person's face. There is a barcode scanner and a currency detection feature that provides information about objects with barcodes and banknotes. Additionally, there is a feature for detecting light in the scene which turns the light intensity into sounds with different frequencies. It is an intuitive method for light detection in the environment. A color detection feature (FR 6) is also included, but it has a poor performance in many cases according to our tests. Obstacle detection and object location/tracking (FR 5,7 and 9) are also not considered in this solution probably due to the hardware limitations of the smartphones. The text detector and reader are one of the power points of this app. There are options for "short text", "document" and "handwriting" that are included in the app. Table 12 summerizes the assessment of functional requirements.

Functional requirement	Evaluation
FR 1: The system must have a command interpreter that allows the user to choose between the options provided in the system.	Using a screen reader (iOS VoiceOver)
FR 2: The system must obtain scene data to allow the user to scan the environment/text through a camera.	Getting input from the phone's camera
FR 3: The system must recognize the scene type.	Providing a very general description (e.g. "This looks like a kitchen")
FR 4: The system must describe the objects that exist in the scene.	Providing brief information about few objects in the scene and detecting persons and faces
FR 5: The system must detect the desired object and locate its position in the environment.	
FR 6: The system must detect the color of a desired object in the environment.	Providing color of the center of the frame when user points the camera at any direction
FR 7: The system must detect the obstacles/hazards that are on the user's path.	
FR 8: The system must let the user scan text and know the written words.	Providing options for "short text", "document" and "handwriting"
FR 9: The system must continuously track object locations as the user moves through the environment.	
FR 10: The Text detector module must detect the presence of text in the scene.	By pointing the camera towards a text it will start reading it if there is any.

FR 11: The system must provide feedback based on user preferences (volume, speed, units, etc.) via audio and/or haptics based on user needs.	VoiceOver in iOS provides customization for volume speed of the speech and other options related to the text size and so on.
Additional functionality 1: light detection	Turning the light intensity into sounds with different frequencies
Additional functionality 2: barcode reader	By pointing the camera towards a barcode, it provides the product information.
Additional functionality 3: currency detection	By pointing the camera towards any currency it provides the information.

Table 12 - Assessment of functional requirements in Seeing AI

6.4.2 Non-functional requirements

On the other hand, regarding the non-functional requirements, it has good portability (NFR 2), power consumption (NFR 6), and affordability (NFR 4) since it runs on a smartphone. However, when it comes to latency (NFR 3) and robustness (NFR 1) in the "scene" feature which describes the scene, there is room for improvement. This is because the device must be connected to the internet for that functionality to work. When it is connected to the internet, sometimes it takes some seconds to process the image and return the description of it. However, this helps the solution to not have a very high computational complexity (NFR 7). It runs smoothly on smartphones since most of the tasks are handled on the cloud. In terms of accuracy (NFR 5) there are no quantitative data on the performance of the system but according to the limited tests that we performed, the text detection, currency detection and light detection work very well, but the face detector and the scene descriptor need improvements. Regarding the NFR 8, the security of this app could be compromised since the data taken from the user's camera is sent to the Microsoft APIs. It is true that Microsoft is a known service provider, but still, exposing users' data to online services can be a privacy and security concern, as it may increase the risk of unauthorized access, data breaches, and misuse by third parties. Moreover,

the solution has some options for personalization (NFR 9). The user can change the pace of the text to speech and the sound of the voice actor. It is also possible to use iOS Siri to interact with the app through voice. But more options regarding haptic feedback could be added to the app for some users such as the deafblind community. Table 13 summerizes the assessment of non-functional requirements.

#	Non-functional requirement	Evaluation
1	Robustness	The system works in different scene dynamics however, when there is no internet only the text recognition works
2	Portability	Runs on smartphone which is a suitable portable device
3	Latency	The text detection is works very fast bu the scene description takes around 2 to 10 seconds for each request
4	Affordability	It is affordable since it runs on various smartphones
5	Accuracy	No quantitative data on the performance of the system
6	Power consumption	Runs on smartphone battery and this could be considered as low consumption. However, it was not possible to measure the exact power usage of it.
7	Computational complexity	Runs smoothly on smartphones since most of the tasks are handled on the cloud.

8	Security	The data is sent to Microsoft servers for processing. This can rise some security and privacy.
9	Personalization	Possible to set the pace of the text to speech and the sound of the voice actor. Also compatible with Siri on iOS.

Table 13 - Assessment of non-functional requirements in Seeing AI

6.4.3 User testing and feedback

In our user research, 4 of the participants mentioned having used Seeing AI for various tasks. For instance, in the case of FR 3 (scene description), P#2 and P#7 mentioned that this functionality still needs improvements to satisfy their expectations. In contrast, the barcode reader and the text recognition functionalities had an acceptable performance from their point of view (P#2,3,5,7). Additionally, in [179] the performance of the text detection (OCR) functionality of this app was evaluated with 7 participants with visual impairment. They used text with differing characteristics such as print size, contrast, and light level in their evaluation. The participants managed to complete 71% of the tasks.

In another study [180] Seeing AI was used for image description, and it was compared with two other solutions (ImageExplore and Facebook image exploration). The solutions were tested with 12 blind participants. Despite the usefulness of the feature, Seeing AI had a lower performance in comparison with the two other options which were better in terms of ease of use and information presentation.

It is important to note that our framework might cover only some aspects of a specific assistive solution which could be designed for additional use-cases not included in our study. For example, in the case of Seeing AI it is clear that the focus of the developers was not on navigation and obstacle detection. Instead, they focused more on features such as barcode reader, currency detection and text detection. Consequently, currency detection, light detection and face detection sub modules can be added to the object detection module in the reference architecture. Our proposed framework is a good starting point for the designers and developers that want to develop an assistive solution for the blind and it has the potential and flexibility to

be extended and/or modified according to the different use-cases and intended user groups. However, it is always important to consider the intended context of use before applying the framework to a specific solution.

6.4.4 System architecture and technological approach

Although the exact architecture of the system could not be fully determined, but interacting with the app revealed that the key functional components appear to be scene scanner, person detector, light detector, color identifier, text reader and currency detector. The exact technologies used in the solution could not be clearly identified. However, for the scene description, it seems to be using an API similar to the one we mentioned in Table 11 that uses Google Vision API. Microsoft has a similar API for the same purpose that could be possibly utilized for this solution. Regarding the text reader, it has two modes, one for short text and the other one for reading documents. The first one seems to be using a local OCR model similar to the one we mentioned in Table 11 (Google Tesseract). However, the latter appears to be using an online model because it does not work when there is no internet connection. The technology behind the other modules (color identifier, light detector and currency detector, barcode reader) were not possible to be identified, but they all operated without requiring an internet connection, except for the barcode reader.

7 Proposed solutions

According to the framework proposed in the previous chapter, two solutions are presented here to fulfil the objectives of the thesis. One solution is focused on exploring the potential of different technologies for the implementation of a cost-efficient assistive solution and the second one is an endeavour to implement a commercialized solution which could be accessed on personal smartphones of the end users.

7.1 Solution 1 – portable laptop version (AssistDiv)

To explore the potentials of various technologies for developing a solution for the P-VI/blindness in indoor environments, it was decided to build a version of the solution on a powerful laptop that utilized an external potent RGB-D camera to reduce the technical limitations of implementation. First, use cases for scene understanding and object location were selected to customize the general architecture to this solution.

Then the process of technology selection was undertaken considering the functional and non-functional requirements. Various object detection algorithms and depth detection methods were compared to figure out which one is more suitable for the intended purpose. In the following sections each step is discussed.

7.1.1 Use cases and architecture

The use cases of the solution were selected in way to address the most important needs of the P-VI/blindness regarding the scene understanding and object location tasks. For scene understanding, we decided to choose Use case 2 (scene description with static objects and standing user) and Use case 7 (scan items on a table with static objects and sitting user).

For object location use cases we chose Use case 11 (finding a small far static object with standing user) and Use case 15 (finding a small static object with sitting user and static objects on a table). The reason for choosing the mentioned use cases was to start with tasks that have less complications for the implementation and testing. Once we had the use cases defined, the architecture of the system was adapted based on the general architecture mentioned previously.

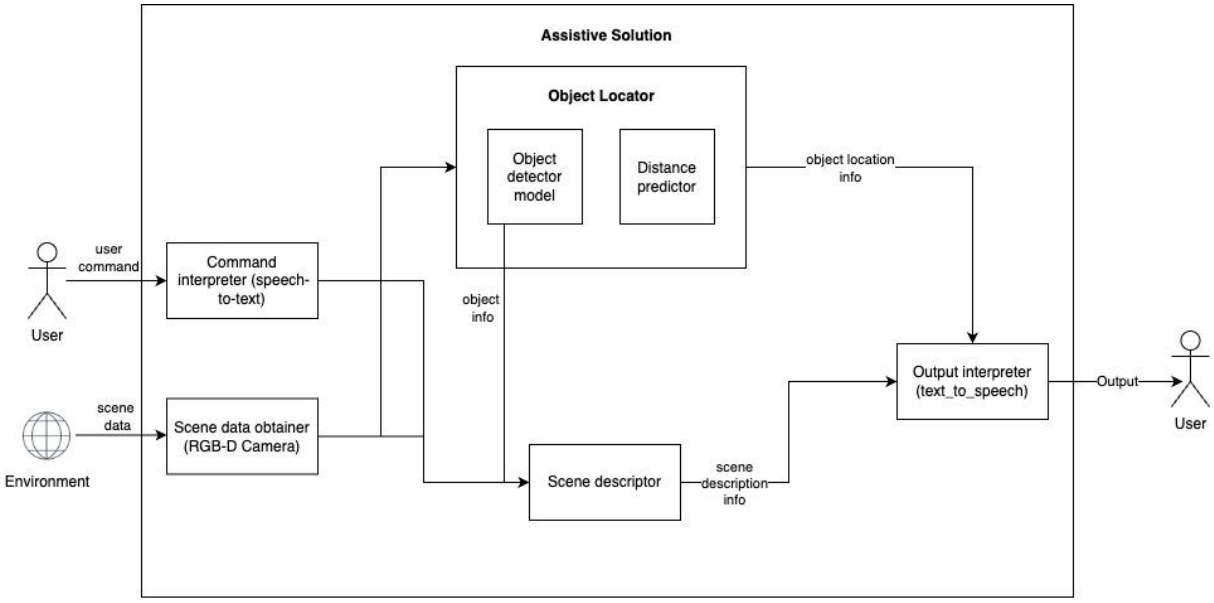

Figure 11 - AssistDiv Architecture

7.1.2 Technology selection

To select the right technology for both hardware and software of the system, a series of evaluations were done to move forward with the approaches that fits us the best. As part of evaluating various object detection methods, one of the tasks was the evaluation of different YOLO versions and exploring the possibility of using transfer learning for adapting it to different use cases. In [181], transfer learning techniques were used to improve YOLOv5 model by retraining them with different configurations and datasets. The goal was to train a model for scenarios with specific objects that provides a good balance between speed and precision, making it suitable for real-time assistance of the P-VI/blindness. The first part of the work involved a thorough analysis of various object detection models, comparing their performance in terms of speed and accuracy to identify a suitable model for real-time applications on devices like the Raspberry Pi 4. The second part of the research delved into transfer learning, where pre-trained models were further trained on specific datasets to enhance their performance for the intended use case. The results of the experiments indicated that while transfer learning can significantly improve model performance, especially in scenarios with limited data, the effectiveness of this approach is heavily dependent on the choice of model, dataset, and training configuration.

The results of transfer learning on YOLO5 were as follows:

Initial Performance Issues: In YOLO and similar models, each bounding box prediction includes an "objectness score". This score is a probability value between 0 and 1, indicating how likely it is that the region defined by the bounding box actually contains an object, as opposed to empty space or irrelevant background. When transfer learning was first applied to the YOLOv5 model using a subset of the COCO dataset [182] and frozen layers, the results were disappointing. The model struggled with poor performance (mAP in experiments was 0.080), largely due to issues related to hard augmentation settings and an objectness loss that failed to decrease properly during training. This indicated that the model's ability to accurately detect objects was compromised, as seen in experiments with the VOC dataset [31]. However, with unfrozen layers the results regarding mAP improved up to 0.32.

Improvements with the OIDataset [183]**:** Subsequent experiments using the OID dataset showed better results in the case of having frozen layers in the training process, suggesting that the choice of dataset and adjustments to the training configuration had an impact on the model's performance. The model's accuracy and detection capabilities improved (mAP 0.16 with frozen layers) in comparison with the previous dataset when appropriate modifications were made to the transfer learning process, such as tuning hyperparameters and adjusting the layers being fine-tuned. However, the overall results were not higher than 0.18 mAP in this case which had room for improvement.

After rounds of testing with YOLOv5 and trying transfer learning techniques, eventually, we decided to move forward with a Mask R-CNN [41] instance segmentation model for detecting objects. This choice was done according to the technology selection criteria (Table 11). This is because it makes the process of depth detection easier since it provides the exact pixels of each object besides the object type. More details regarding the instance segmentation method are provided in the next sections.

For the depth detection, we utilized Intel Realsense D455 RGB-D camera because it provides real-time high precision depth map of the images. We evaluated this camera against another high-end depth camera (Azure Kinect DK) because of its superior performance. To evaluate the two cameras, the accuracy of their depth detection module with various objects (such as bottles, tea boxes, white boxes, tea boxes, and patches) that had varying shapes and colors under different lighting conditions was compared. Realsense camera had a lower average error rate (-10.2 millimeters) in comparison with (51.4 millimeters). The full comparison is in Table 14.

Object	Camera	Ground truth (GT) short range (mm)	Short range (mm)	GT mid range (mm)	Mid range (mm)	Mid range with low lighting (mm)	GT long range (mm)	Long range (mm)	GT AVG (mm)	AVG (mm)	GT-Predicted Error (mm)
Matches	Azure	605	583	1270	1192	1196	2010	1996	1295	1257	38
Matches	Realsense	605	590	1270	1205	1222	2010	2050	1295	1282	13
Tea box	Azure	595	582	1210	1190	1193	2005	1985	1270	1252	18
Tea box	Realsense	595	602	1210	1223	1236	2005	2065	1270	1297	-27
Mouse box	Azure	630	578	1230	1185	1183	2015	1985	1292	1249	43
White box	Realsense	630	595	1230	1207	1197	2015	2230	1292	1344	-52
White box	Azure	680	580	1270	1207	1200	2045	2000	1332	1262	70
Bottle	Realsense	680	570	1270	1197	1205	2045	2230	1332	1332	0
Bottle	Azure	700	570	1255	1177	1187	2040	1985	1332	1244	88
	Realsense	700	639	1255	1255	1245	2040	2057	1332	1317	15
RealSense Avg Err. low light							-26 mm	**RealSense Avg Err.**			-10.2 mm
Azure Avg Err. low light							-55.2 mm	**Azure Avg Err.**			51.4 mm

Table 14 - Camera comparison

We also compared the performance of deep learning methods for depth detection such as MiDaS [184] but there was a huge performance gap (13.43% mean absolute relative error (MARE) which is predicted depth versus ground truth depth for MiDAS v2.1 [171]) between RGB-D cameras (millimeters in case of D455 [185]) and deep learning techniques. As a result, we chose the D455 camera due to its superior real-time performance, as RGB-D cameras can detect depth in milliseconds, while MiDaS takes seconds to process each frame. Additionally, RGB-D cameras offer greater consistency, being less affected by environmental variations that can impact deep learning models [186]. Furthermore, using a dedicated depth camera simplifies the overall system architecture compared to the complexities of implementing and adapting a deep learning module. A potent laptop (HP OMEN - 15-dc0030ns) was utilized to run the system. The depth camera was connected to the laptop running the instance segmentation algorithm. The camera was installed on the user's head using a camera headband. A Bluetooth neckband (QCY-C1) was connected to the laptop that enabled the user to interact with the solution through voice and hear the system's feedback through headphones (Figure 12).

Figure 12 - A user wearing the solution including the backpack with the laptop, camera and the headphones

7.1.3 Workflow

The flow chart of the system's workflow is in Figure 13. The user chose the desired options using a voice user interface that worked using voice commands through the Bluetooth neckband.

When the system started, it asked the user to choose a preferred language by saying "English" or "Spanish". Once the language was chosen, there were two main options, one for scene understanding and the other one for finding objects.

7.1.3.1 Scan the scene

For scanning the scene, once then interaction language is chosen, the system asks the user to choose the desired option (locating objects or scanning the scene). Once the user says, "Scan" the system then asks if the user wants a general description, or a detailed description. The user can reply by saying "detailed" or "general". The general description is based on the category

of objects. For example, a microwave, a toaster and a pan are included in the category of appliances. It provides the category of the objects detected from left to right side of the screen. For example, “Objects are accessory, furniture, food”. The detailed description also names the objects from left to right side of the scene describing their distance from the user. For this purpose, the image captured from the scene is divided into three parts (left, front and right) and the scene description informs the location of the objects accordingly. For instance, “objects on your left side are, keyboard at 2 meters, chair at 1 meter, objects in front of you are cell phone at 1.5 meters, person at 2.5 meters” and so on.

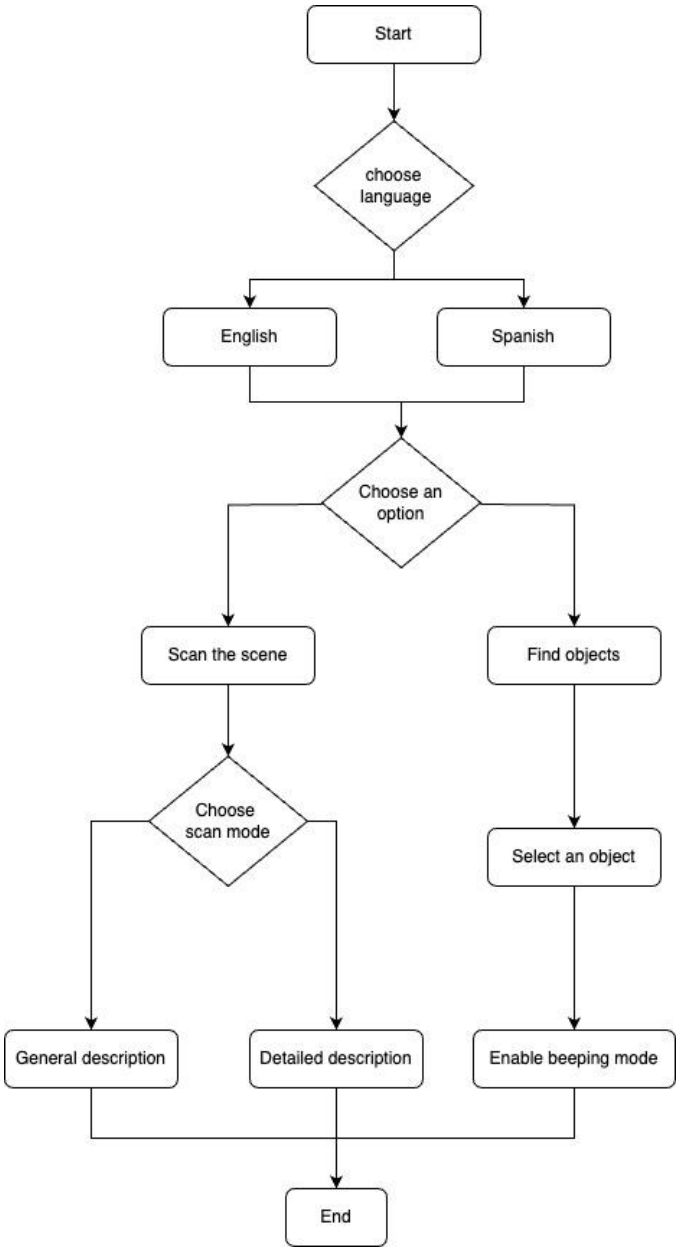

Figure 13 – AssistDiv flow chart

7.1.3.2 Locate objects

To locate an object, the user should choose the option for finding objects instead of scanning the scene by saying “find objects”. Then the system provides a list of objects present in the scene. Then the user can choose an object and the system provides information regarding the distance and location of that object (e.g. book is on your right side 2 meters away). Afterwards, the user has the option to try to reach it using the beeping mode. Once the beeping mode is enabled, the system beeps when the object is in the center of the frame. If not, it informs the user if it is on the right or left side of the scene so that the usar can turn up to the point when the object gets in the center of the screen. As the user gets closer to the object, the beeping mode frequency increases. Once the distance of the user from the object is less than one meter, the system tells the user to “stop”.

7.1.4 Technical information

For the development of the solution, the Python programming language is utilized along with the frameworks needed. Detectron2 [165] by Meta, a platform for object detection, segmentation and other visual recognition tasks is used to run the instance segmentation algorithm. Mask R-CNN [41] model with a ResNet-50-FPN backbone instance segmentation model which was trained on COCO dataset is used for the object detection task. Moreover, for the depth detection the Intel RealSense cameras have a framework (PyRealsense) for python which facilitated the process of getting the depth map of the environment and processing the results.

When user activates the "scan the scene" option, the camera takes one image containing the depth map of the environment and an RGB image which is sent to the instance segmentation algorithm. Figure 14 shows an example image of "scan the scene" mode showing the objects detected in the scene and their distance from the camera in meters. Depending on the particular "scan the scene" mode (general or detailed) the system provides a description through the voice user interface which is discussed in more detail later in the user interface section.

Figure 14 - Screen shot of the solution capturing the instances and their distances in the scene

The object categories are as follows, with the objects included in each one:

"person": ["person"],

"vehicle": ["bicycle", "car", "motorcycle", "airplane", "bus", "train", "truck", "boat"],

"outdoor objects": ["traffic light", "fire hydrant", "stop sign", "parking meter", "bench"],

"animal": ["bird", "cat", "dog", "horse", "sheep", "cow", "elephant", "bear", "zebra", "giraffe"],

"accessory": ["backpack", "umbrella", "handbag", "tie", "suitcase"],

"sports": ["frisbee", "skis", "snowboard", "sports ball", "kite", "baseball bat", "baseball glove", "skateboard", "surfboard", "tennis racket"],

"kitchen": ["bottle", "wine glass", "cup", "fork", "knife", "spoon", "bowl"],

"food": ["banana", "apple", "sandwich", "orange", "broccoli", "carrot", "hot dog", "pizza", "donut", "cake"],

"furniture": ["chair", "couch", "potted plant", "bed", "dining table", "toilet"],

"electronic": ["tv", "laptop", "mouse", "remote", "keyboard", "cell phone"],

"appliance": ["microwave", "oven", "toaster", "sink", "refrigerator"],

"indoor objects": ["book", "clock", "vase", "scissors", "teddy bear", "hair drier", "toothbrush"]

The pseudo code for “general” and “detailed” mode is as follows:

If mode is 'detailed':

for each object in detected_objects:

Set text to the name of the object

for each object in detected_objects:

get all x coordinates of the pixels occupied by the object (x_pixels)

calculate left_pixels as the number of pixels with x coordinates less than left_threshold

calculate right_pixels as the number of pixels with x coordinates greater than right_threshold

calculate front_pixels as the number of pixels with x coordinates between left_threshold and right_threshold

if left_pixels is greater than both right_pixels and front_pixels:

add object to left_objects list

else if right_pixels is greater than both left_pixels and front_pixels:

add object to right_objects list

else:

add object to front_objects list

if left_objects is not empty:

call speak with message "objects on the left side are:" and the language

for each object in left_objects:

call speak with object's name and distance in meters and the language

if front_objects is not empty:

call speak with message "objects in front of you are:" and the language

for each object in front_objects:

call speak with object's name and distance in meters and the language

if right_objects is not empty:

call speak with message "objects on your right side are:" and the language

for each object in right_objects:

call speak with object's name and distance in meters and the language

Else if mode is 'general':

Create a set of distinct categories from detected_objects

For each category in the set:

Set text to the category

Speak the text in the specified language

Moreover, the object location mode uses the average depth points of each object's mask to estimate their distance from the camera. When the beeping mode is enabled to reach an object, this process happens in every frame. However, since the instance segmentation model takes around 2-3 seconds to process the image, the beeping sound is not played instantaneously. This makes the beeping mode a little bit confusing for the users which is discussed more in the user testing section. Additionally, the object location module has another limitation. If more than one instance of the object exists in the scene, it is not possible to distinguish between them in every frame. For example, if there are two books in the scene, moving the camera and detecting the books in several frames will change the detection order which makes it impossible to distinguish between them. This is because the instance segmentation algorithm is unable to differentiate between two books in each frame and keep this information. We tried to make a distinction by adding a numbering to the objects of the same type (e.g. book 1 and book 2) but the indexing would differ in every frame since the model does not have a memory of previous frames. Therefore, in the current implementation there is a limitation of only locating object categories when only one instance of them exists in the scene.

All of the programming codes regarding the AssistDiv can be found open source on its Github repository [187].

7.1.5 User interface

As it was mentioned previously, the user interface used for the system is a voice user interface. Initially, Google speech recognition is used for the user speech detection as the input of the system. The SpeechRecognition [188] python framework is used for that purpose. However, since the response time of the Google's API would differ from time to time and sometimes it was very slow (each request took around 30 seconds in some cases) it was decided to utilize an offline model for speech recognition. Vosk Speech Recognition Toolkit [189] is used for that purpose. Small models (for higher inference time) of English and Spanish language are used to detect the user's input in both languages. For the output, Google's gTTS API [190] is used for text-to-speech which provides a human-like voice.

The process of getting the user's command is a challenge since the speech recognition does not work perfectly if the user does not pronounce the word with American accent or in the case of Spanish language, Madrilenian accent. Additionally, in some cases user wouldn't say the command correctly (e.g. saying "scan the scene" instead of just "scan") which made the interaction problematic. To tackle this problem a Fuzzy string-matching tool is used called RapidFuzz [191]. This library provides a fuzzy ratio based on the similarity of two strings. If the user's input is more than 50% similar to what the original input command is, then it will be accepted. For instance, if the user says "detail" instead of "detailed" which is the correct command, the system still accepts it.

7.2 Solution 2 – smartphone version (V-ASSISTANT)

After the development of the laptop version and exploring different possibilities for the development of the solution, it was decided to make another version of it to assess the potential for the industrialization and commercialization of the work. It is important to note that the development of this solution is still an ongoing process. Since it is being developed for commercial purposes, the source of the project is not open source at this moment.

The process of the development of this version started by participating in a competition. In 2022, I had the chance to participate in ACTUA UPM 19 and present the idea of this solution in the competition. It was selected as one of the final candidates of the competition which gave me the opportunity to learn how to turn an idea into an actionable business plan. During some workshops I learned how to pitch my idea and turn it into a commercialized product. A year later, I took part in the EIT Digital Venture Program 2023. Applying the skills I had learned

from my prior experience, I presented the idea again and it was selected as one of the top ideas chosen in Spain and received some funding for the creation of a minimum viable product (MVP). Therefore, the development approach for V-ASSISTANT is directed more toward commercialization and developing a solution for the intended users.

7.2.1 Competitive analysis

One of the major steps of developing any commercial product is knowing the pros and cons of the existing solutions and identifying a competitive advantage. For this reason, first we analyzed the existing solutions and then thought of our own competitive advantage which is object location and offline scene description. In the competitive analysis, five different popular assistive solutions were assessed. Seeing AI [102], Envision AI [102], Lookout [103], Be My Eyes [105], Super Lidar [192] with emphasis on their strengths, weaknesses, scene description capabilities, and depth estimation features.

Microsoft's Seeing AI is known for its brand reputation and ability to detect color, text, barcodes, and currency; however, it falls short in providing detailed object descriptions within a scene and lacks depth estimation capabilities. Envision AI includes smart glasses that provide scene description, but the high cost of the glasses (2000-3500€) is a significant disadvantage. Similarly, Google's Lookout is effective in object and text detection but is only available on Android and does not provide depth estimation. Be My Eyes has a large user base, but it lacks an offline mode and does not support depth estimation. Super Lidar, on the other hand, is designed for high accuracy depth detection, but it is only compatible with iPhones equipped with LiDAR sensors.

Our solution’s competitive advantage is in distance estimation, and voice user interface. Additionally, V-ASSITANT provides an offline model for scene description and object location which is another competitive advantage for the situations where user does not have access to the internet or has privacy concerns. Table 15 includes the detailed advantages and disadvantages of each solution.

	Strength	Weakness	Scene description	Depth estimation
Seeing AI	Brand reputation (Microsoft) Color, text, barcode, and currency detector	Describes the scene, not each object.	✓	✗
Envision AI	Comes with a glassware	Glassware is expensive (2000-3500€)	✓	✗
Lookout	Reputación de marca (Google) Precisión en detección de imagen/texto	Solo para Android	✓	✗
Be my eyes	Large clientele	Without offline mode	✓	✗
Super Lidar	High accuracy for depth detection	Only works on iPhones with LiDAR sensor, Not very user friendly	✗	✓
V-ASSISTANT	Distance detection and tracking mode, Voice user interface, Offline object detection model	Only Android version (for now)	✓	✓

Table 15 - Competitive Analysis

7.2.2 Use cases and architecture

Similar to AssistDiv, two main features of the smartphone version are scene understanding and object location. In order to design the architecture of the solution, the same use cases related to object location and scene understanding as AssistiDiv were considered. However, we addressed the use cases differently in the process of the design and development of this solution since we were competing with similar solutions that provide AI powered features. Figure 15 shows the architecture of the solution. The main difference in the architecture is in the way that the user interacts with the system which is the screen reader (Android's Talkback [193]) and the speech-to-text module.

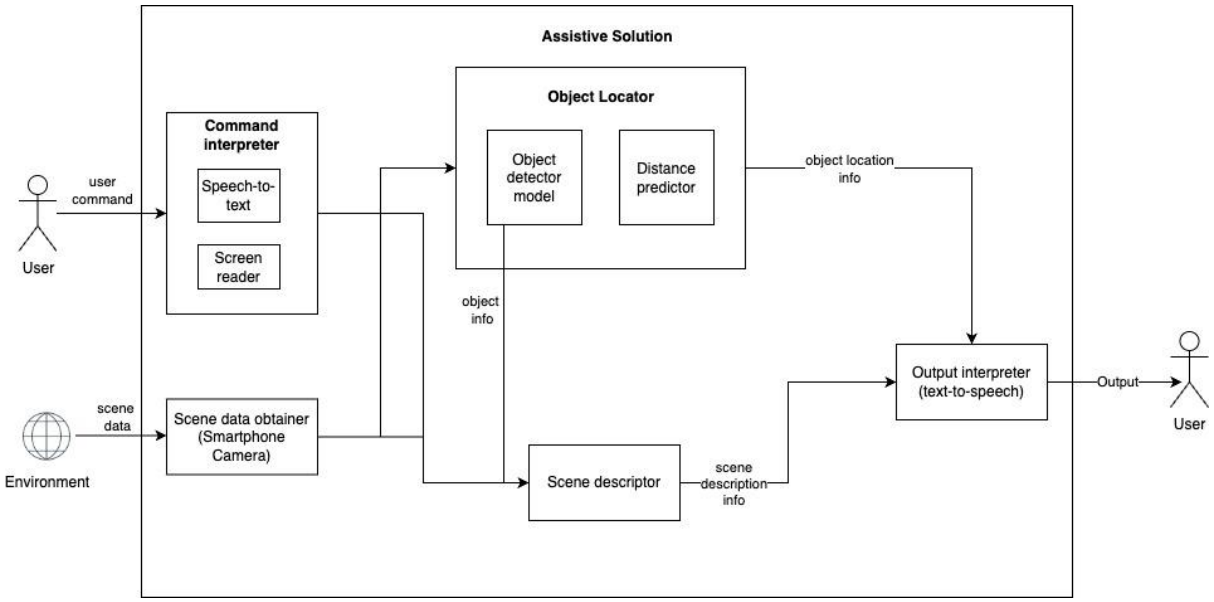

Figure 15 – V-ASSITANT architecture

7.2.3 Technology selection

Taking into account both functional and non-functional requirements, the technology selection process was started. Since we had limitations regarding the hardware (Android smartphone) of the solution, it was not possible to use potent instance segmentation algorithms for detecting objects or a depth camera to calculate the exact distance of the objects in the scene. For this reason, other approaches were considered in the design process.

The device used for the implementation of the solution had Android operating system since it is open source and provides a variety of frameworks for computer vision tasks. Additionally, android phones have more reasonable price range specially outside the US.

The feature regarding the scanning the scene and telling user about the existing objects is implemented using two approaches. The first approach is using a cloud API for describing the objects in the scene (Figure 24). We used a multimodal large language model (MLLM) API (Anthropic's claude-3-opus model [194]) which became popular recently in similar commercial solutions such as Envision [102] or Be My Eyes [105]. However, since these models need continuous connection to the internet, we also considered an offline object detection model that can provide the list of objects present in the scene when user does not have access to the internet. The object location mode is pretty similar to the AssistDiv. It first provides a list of objects by capturing a photo of the scene including their location. The depth

detection is handled using Google AR Core which provides the depth map of an image using's its Depth API. It generates depth images using a depth-from-motion algorithm, providing a three-dimensional view of the world. Each pixel in a depth image represents the distance between the camera and the scene. This algorithm compares multiple device images from different angles to estimate the distance between each pixel as a user moves their phone. It selectively employs machine learning to improve depth processing, even with minimal user input. It also takes advantage of any additional hardware that a user's device may contain. If the device includes a dedicated depth sensor, such as time of flight (ToF), the algorithm will automatically combine data from all available sources. This improves the existing depth image and enables depth even when the camera is not moving[195].

7.2.4 Workflow

Despite the similarities in use cases between the two solutions, V-ASSISTANT has considerable differences with AssistDiv in terms of its interaction design. In order to use the solution, the users have to login in the app using their Google accounts. The user login/registration is considered to track the user activities and the possibility to monetize the solution in the future. To make the registration process easier for the P-VI/blindness, we considered the registration only with the google account option. This way the user can click on one button and enter the app (Figure 16) without having to fill any entries.

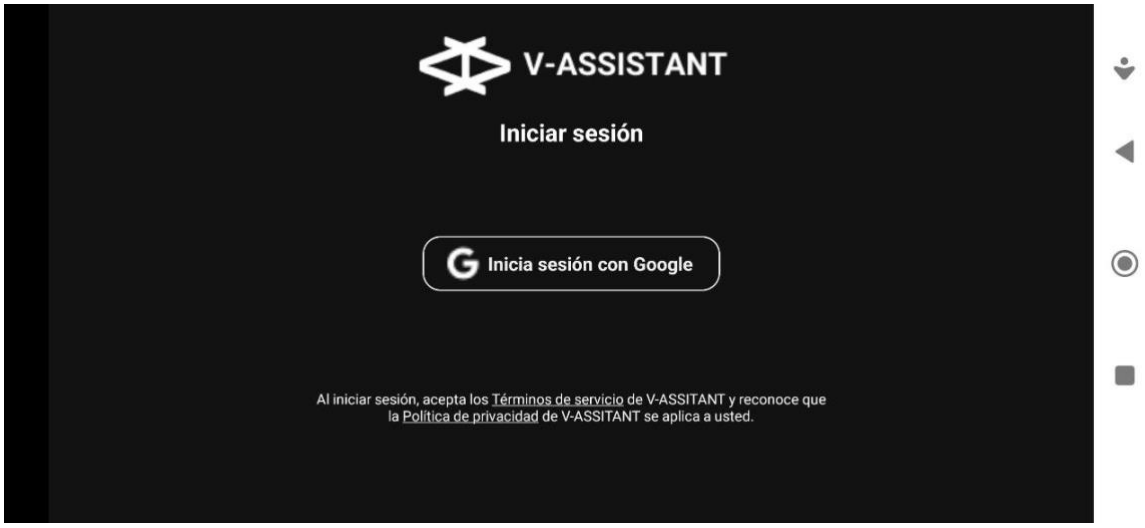

Figure 16 – App's user registration

Like the previous solution, the smartphone version also operates in both Spanish and English languages. The user can choose the preferred language from the main menu of the app (Figure 17).

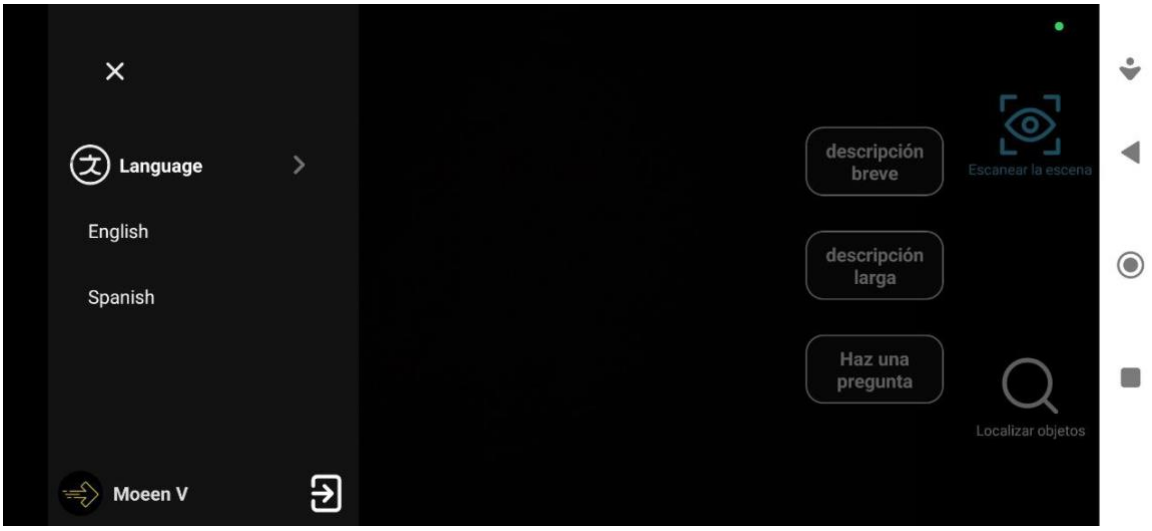

Figure 17 - App's language selection

Once the user logs in the app, the two main features of scanning the scene and locating objects can be accessed. These two main features are discussed in the next sections.

7.2.4.1 Scan the scene

As mentioned previously, scanning the scene can be achieved in two online and offline modes. When the device is connected to the internet, the Anthropic's API is used to infer the objects in the image. There are three options for the user to select, which are: "short description", "long description" and "ask question". The first scene information displayed in Figure 18 is the short description mode that provides a brief description of the scene, and the second one is the description provided by the "long description" mode, which provides a more detailed description.

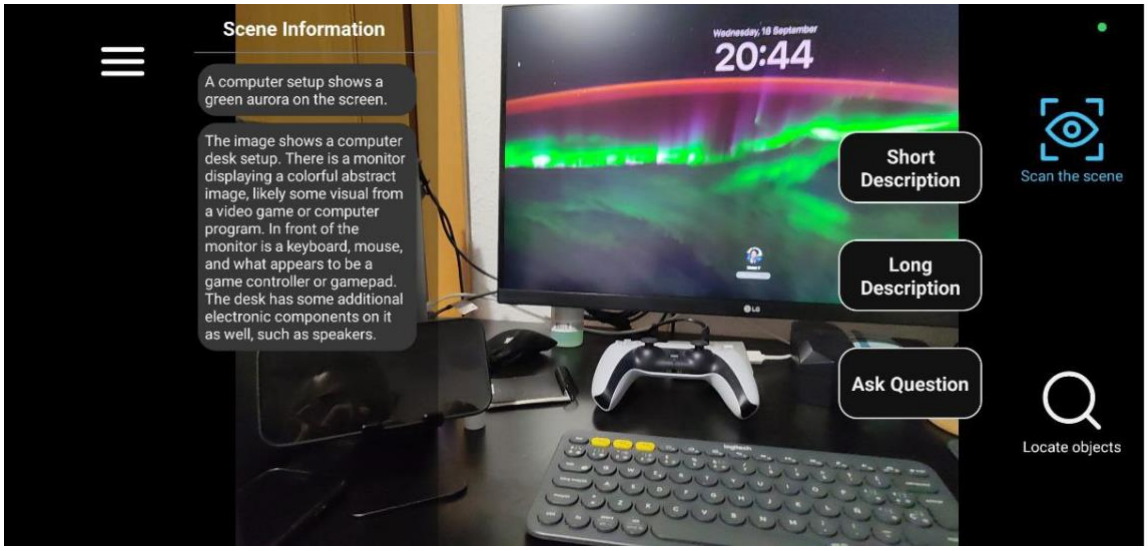

Figure 18 - Scan the scene mode (short and long description)

The “ask question” button allows the user to ask a question about the surrounding environment through speech. Figure 19 shows the question asked by the user and the description provided by the system.

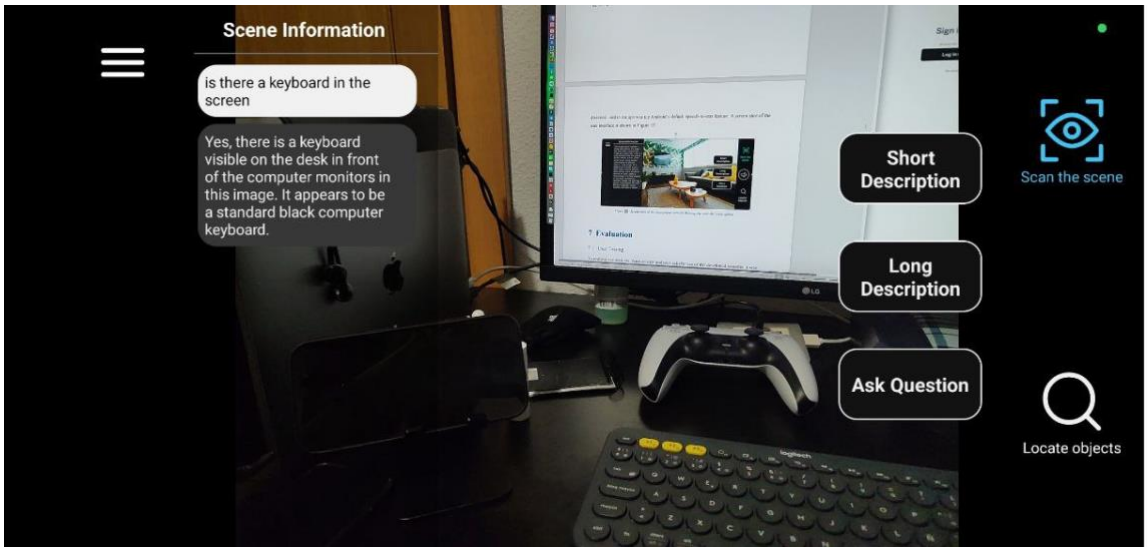

Figure 19 - Scan the scene mode (Ask question)

When the device does not have access to the internet, the offline model (EfficientDet Lite4 V2 [196]) on the device starts to work as an alternative. However, it only provides a list of objects (Figure 20) with limited explanation (e.g. “4 objects detected in the scene. From left to right side of the frame objects are: mouse, keyboard etc.”). The offline model is the default version of EfficientDet Lite4 V2 trained on COCO dataset which covers 80 different object classes.

Figure 20 – V-ASSISTANs offline scene description

7.2.4.2 Locate objects

Locating objects feature has a similar interface to scan the scene mode. The user can capture an image and then the system provides the list of the objects using the offline model (Figure 21).

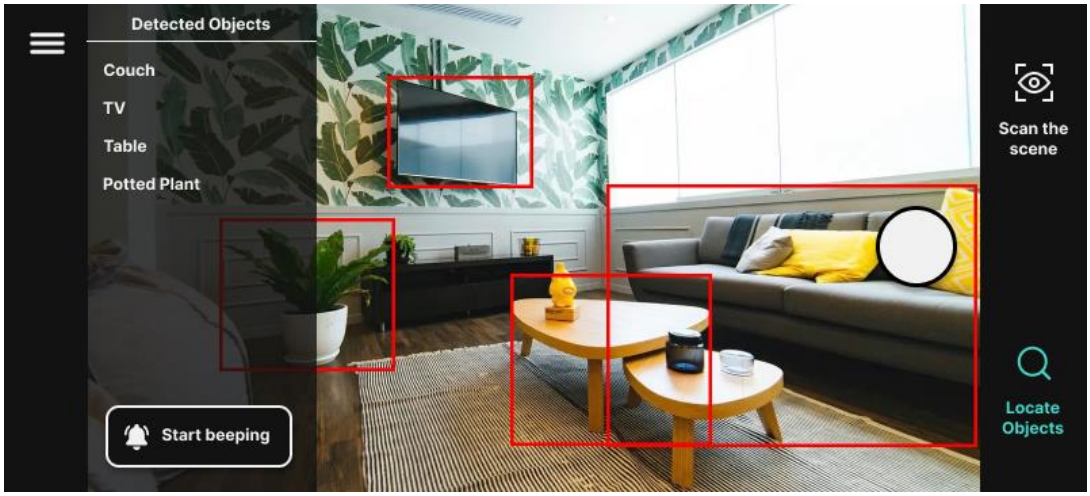

Figure 21 - Locate objects mode

The cloud API could not be used for this mode because it only provided the description of the image and not the position of the objects in the image in a way that could be used along with the distance detection module. This is because the coordinates of the exact pixels of objects is needed for estimating the distance. Once the user captures an image and the system provides the list of objects, by tapping on each of them, the system informs the user about the location of that specific object (e.g. couch is on the right side of the screen 2 meters way.). When there

are more than one instance of an object category, for example “two books”, it provides 2 books in the list instead of book 1 and book 2 which is the approach in AssistDiv. However, it will not provide the location of the two books if user taps on it. This feature is planned to be added in the future.

Regarding the beeping mode, once the user selects an object and enables the beeping mode, the frequency of the beeping mode gets more frequent if user gets closer to the object and the object is located in the center of the screen. However, at the point of writing this, there were some challenges regarding this feature. Google AR Core’s depth API is not included on all Android phones and only newer ones include this feature. Additionally, this depth API cannot always provide the right estimation since it uses machine learning for depth processing which can increase the error rate. Additionally, it works with some delay (2-3 seconds of delay or even more in phones with weaker processors) which makes the beeping mode less effective. Besides, this API has technical issues that sometimes cause crashes which lead to the closing of the app. For this reason, depth estimation was only included in the preliminary versions and has been removed temporarily from the final beta version to not impact user experience.

The flow chart of the system is shown in Figure 22.

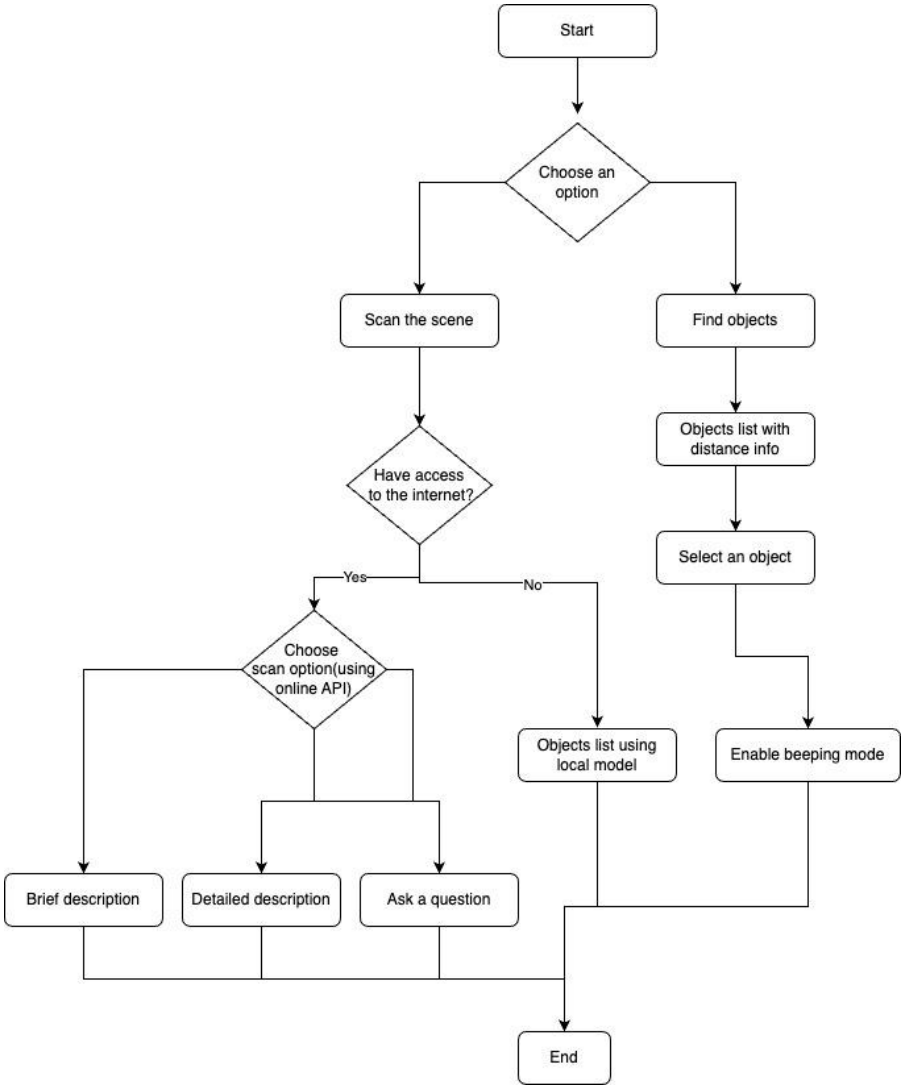

Figure 22 – V-ASSISTANT flow chart

7.2.5 User interface

The design of the user interface was first approached by analyzing other similar solutions. We decided to design a very simple user interface that had less complications for the P-VI/blindness. This is because using an android device through a screen reader (Talkback) makes the interaction slower and we wanted to make the experience as efficient as possible.

Before designing the high-fidelity version of the user interface, we started with low-fidelity prototypes to develop more concrete ideas about the solution. For example, the app was designed in horizontal mode so that the camera could capture wider photos, resulting in more scene data. We did not have the chance to test our interface with end users at the beginning but,

based on the analysis of similar solutions, we came up with the initial design (Figure 23). However, it changed and elaborated as we moved on with the development.

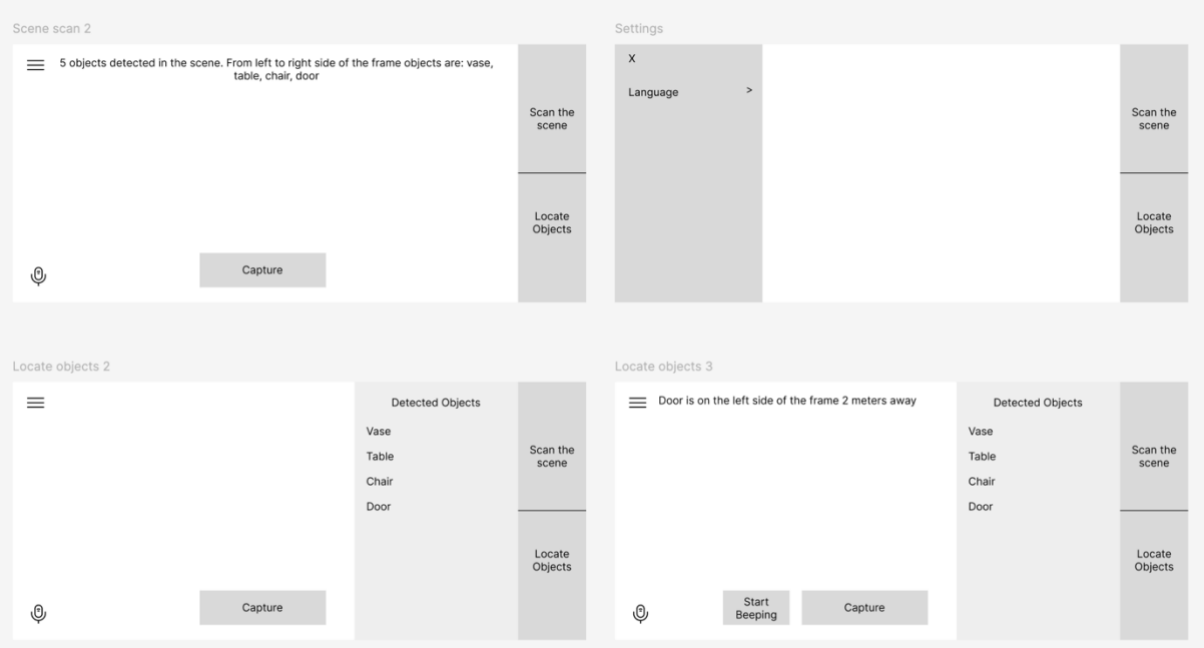

Figure 23 – V-ASSISTANT low-fidelity prototype

The interaction of the user with this solution was different in comparison with solution 1. In the previous solution, the device was wearable and every interaction of the user with the system was through a voice user interface. However, in the case of solution 2, the interaction was possible through the touch screen using a screen reader (Talkback) or the speech-to-text feature (e.g. asking questions about the scene using voice). The voice detection used in the app was the Android's default speech-to-text feature (Figure 24).

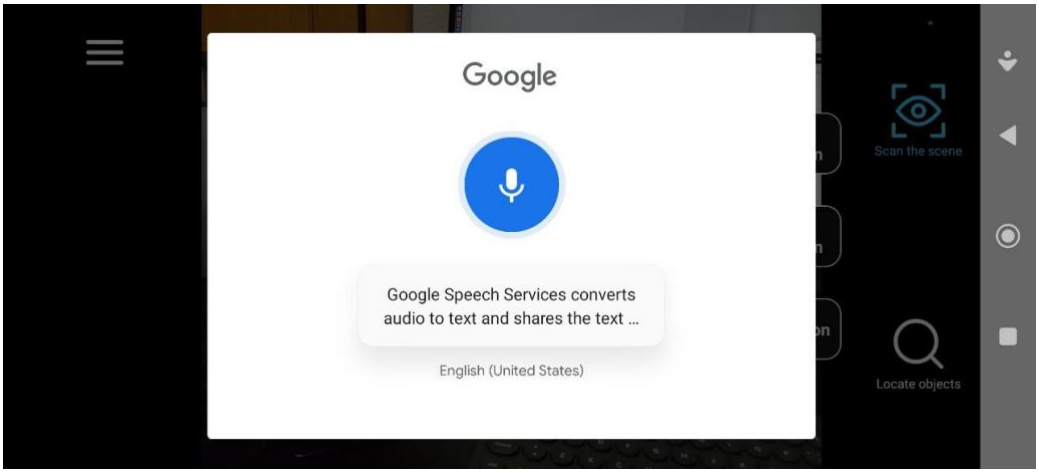

Figure 24 – V-ASSISTANT speech-to-text feature

During the implementation of the user interface, one of the main challenges that we had was the differences in devices. Talkback would work differently depending on the Android's version and smartphone's brand. For example, on one of the testing devices (Xiaomi Redmi Note 12 Pro), Talkback did not work properly when the user wanted to open the menu and change the language. This problem did not exist on Samsung devices. The user interface of this app still has room for improvement and needs additional testing with more end users and on more devices.

8 Evaluation

8.1 User Testing

After the development of the solution (AssistDiv), we decided to assess its usability, functionality and user satisfaction. To accomplish that, it was decided to test the AssistDiv with blind-folded users in a controlled environment. This helped to figure out the advantages and disadvantages of the solution before testing it with the P-VI/blindness. The design of the test scenarios was according to the selected use cases that the solution was addressing. The selection of different objects for each scenario was according to the objects that could be detected by the implemented object detection algorithm dataset. We chose the number of objects present in the scene according to the 'Miller's Law' (7±2 rule) [197] which explains that people can only hold seven plus or minus two items in their short-term memory. Moreover, the questionnaires used in the evaluation were designed to gather qualitative and quantitative data. Once the solution was tested, the evaluation results were used in improving the smartphone version (V-ASSISTANT) to be tested with the intended users.

8.1.1 AssistDiv testing

To begin the testing process, a test plan was prepared and approved by the Ethics Committee of the Universidad Politécnica de Madrid under reference code: CVDASFVU00-ADAJ-HUMANOS-20230619.

The process of testing the solution was undertaken with 6 blindfolded sighted individuals. They were selected among students of the European Master in Software Engineering of Universidad Politécnica de Madrid. The testing was carried on in the facilities of the Madrid HCI Lab [198]. The testing was carried out in two different languages. Three of the testers performed the testing in Spanish language and three of them in English.

First, the purpose of the study, the testing process, and the expected duration were explained to the participants All of the participants read and signed a consent form before participating in the testing. Participants were then allowed to familiarize themselves with the system and its functionalities. A brief tutorial on using the system was provided and any questions they had were answered.

Once familiarized, they participated in three different testing scenarios:

1. Scene understanding
2. Locating objects (short distance)
3. Locating objects (long distance)

All of the tests were video-recorded for further analysis of the users' behavior. Once they finished the scenarios, they answered to a questionnaire about their experience.

In the following sections, each testing scenario and its corresponding results are described:

8.1.1.1 Scene understanding

The objective of this scenario was to examine the capability of users to create a mental map of the scene from the description provided by the system. The system's description output included the position of the objects and their distance from the user starting from the left side (e.g. objects on your left side are: keyboard at 2 meters, cup and 1.5 meters. Objects in front of you are monitor at 1 meter, bottle at 1.5 meters, laptop at 1.9 meters. Objects on your right side are bottle at 1.7 meters, umbrella at 2 meters.).

8.1.1.1.1 Process

1. Set up a table with various items, such as keyboard, bottle, cup, laptop, umbrella, monitor and a chair (Figure 25).
2. Instruct the blindfolded participant to position themselves in the scene.
3. Direct the participant to activate the scene understanding mode and attentively listen to the system's description of the scene. (First in General mode and then in Detailed mode)
4. Ask users to use a drag-and-drop interactive map to position the objects where they were located and compare the user response with the ground truth.
5. In parallel, record observations on the system's usage, success in identifying objects, and any difficulties encountered.

Figure 25 - Set up for scene understanding testing

8.1.1.1.2 Evaluation Instrument

In order to evaluate the results of the tests, we decided to check if testers could form a mental map and identify objects on their right, front and left side with the correct order.

Once the task was completed, the users were asked to use an interactive map to determine the location of the objects presented to them by the system. Seven objects (Keyboard, Bottle, Cup, Laptop, Monitor, Umbrella, Chair) were presented to the user considering the rule of seven. One object (Microwave) presented in the interactive map was not included in the output of the system in order to make the task more challenging. Participants were not informed that one of the objects was not included in the output of the system. This method examineed the accuracy of the user's mental map created after receiving information from the system.

In Figure 26 the interactive map is shown. The smiley face represents the position of the user in the scene and the squares are for determining the position of the possible objects. On the left corner of the screen, circles with the object names are placed. Testers had to take the object

names (circles) and drag them to the corresponding positions (squares) based on the information they received from the system.

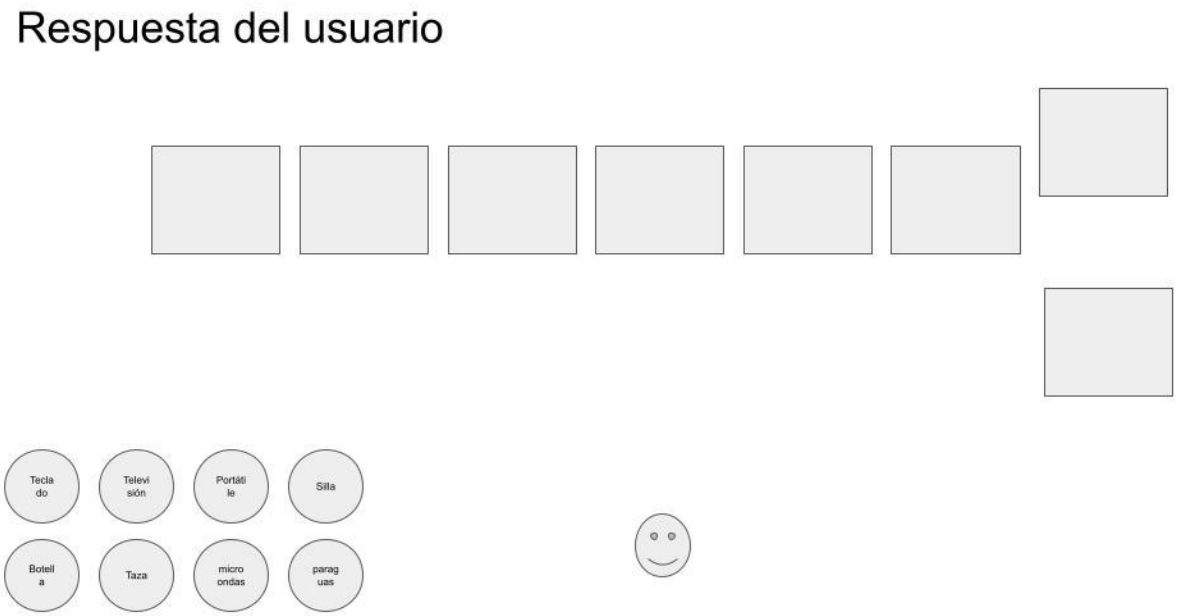

Figure 26 - Interactive map for the evaluation of mental map

8.1.1.1.3 Results

Table 16 shows the number of objects that were correctly positioned on the interactive map by the users. Most of the testers had a better performance identifying the approximate position of the objects (left, front or right) than the exact order of the objects.

As demonstrated in the table, testers could remember the first two objects positioned on the left side better than the other objects. The reason for that might be that objects on the left side were presented to the tester before the objects in front and on the right side. It is important to mention that there was a parameter that affected the system's output. The object detection algorithm did not detect all the seven objects for every participant. This is because depending on any subtle difference of one frame with another one, the results could be different. For example, if the position of the tester, angle of the head or the height of the user were different from the previous one, it could affect the performance of the object detection algorithm and there is a chance that it does not detect a couple of objects.

	Left	Front	Right	Total
Participant#1	0	0	2 (umbrella, chair)	2
Participant#2	2 (keyboard, bottle)	2 (Cup)	1 (umbrella)	5
Participant#3	2 (keyboard, bottle)	1 (laptop)	1 (umbrella)	4
Participant#4	2 (keyboard, bottle)	1 (laptop)	2 (umbrella, chair)	5
Participant#5	1 (bottle)	2 (laptop, TV)	1 (chair)	4
Participant#6	2 (keyboard, bottle)	1 (laptop)	1 (chair)	4

Table 16 - Interactive map results

8.1.1.2 Long-distance object location

The objective of this scenario was to evaluate whether the guidance provided by the system to locate an object far from the user is effective (the user is able to walk towards the object and stop next to it).

8.1.1.2.1 Process

1. Put a specific object (a bottle) in the room within the reach of the area covered by the system's camera.
2. Instruct the participant to utilize the assistive solution to identify an object and its location.
3. Ask the participant to locate the bottle in the room using the beeping mode:
    1. The system first describes the objects and their distances to the tester, then the tester must select the "bottle" from the object list and then activate the beeping mode. If the object is on the center of the screen the beeping would start, and if the object is on the left or right sides of the screen the system would inform that "The selected object is on the left/right side". If the object is outside the image, the system would play a message indicating that "The selected object is not in the frame".

2. Then the participant would start turning and/or walking in order to approach the position of the bottle. When the participant is less than one meter away from the bottle, the system would play the “stop” message.

4. In parallel, record observations on the system's usage, success in identifying and locating objects, and any difficulties encountered.

8.1.1.2.2 Evaluation metric

To evaluate the process of locating objects in long distance, the success of the user in locating the specified object (bottle) was assessed. This means that the user had to get closer to the object until the system provided the “stop” message (Figure 27).

Figure 27 - Blindfolded tester looking for the bottle using beeping mode

8.1.1.2.3 Results

In this testing phase, only 1 out of 6 participants did not manage to locate the bottle in their surroundings. The navigation had a little bit of delay (around 2 seconds) in playing beeps when the bottle was detected, because of the object detection model taking time to process the photos in each frame. For this reason, the participants were asked to move slowly when they were tracking the bottle. The reason for the failure of one of the participants was the fact that another similar bottle was in the scene, and they mistakenly detected that bottle instead of the intended

bottle. This could be considered as a flaw in the testing process. The time taken for the participants to reach the bottle was also measured. On average it took 1 minute and 23 seconds for them to locate the object when they activated the beeping mode. Table 17 shows the time it took for each participant to locate the bottle.

Participant	**P#1**	**P#2**	**P#3**	**P#4**	**P#5**	**P#6**
Time	1 minute (m) 30 seconds (s)	50 s	2 m 10 s	1 m 55 s	Unsuccessful	35 s

Table 17 - Object location time measurement

8.1.1.3 Short-distance object location

The objective of this testing scenario was for to evaluate whether the guidance provided by the system to locate an object close (<1 meter away) to the user is effective to allow the user grabbing the object.

8.1.1.4 Process

1. Arrange various objects on a table (mouse, cup, laptop, bottle, wine glass, and scissors as shown in Figure 28)
2. Instruct the participant to use the system to identify and locate a bottle on the table.
3. Ask the user to enable the find object mode and remember the location of a specific object provided by the system.
4. Ask the user to stand up and point towards the object. This was to avoid potential collisions with the objects in the scene, for the user's safety and to avoid altering the scene from one participant to the next.
5. In parallel, record observations on the system's usage, success in locating objects, and any difficulties encountered.

8.1.1.4.1 Evaluation metric

To evaluate the process of locating an object in short distance, the success of the user in locating the specified object (bottle) was assessed by evaluating the success of the participant in pointing at the right direction towards the bottle.

First, the user had to enable the object location mode and then listen to the descriptions of the system attentively. The system would provide the following description: "Mouse is on your left side at 1 meter, cup is on your left side at 1 meter, laptop is in front you 1 meter, bottle is

on your right side at 1 meter, wine glass is on your right side at 1.4 meter, scissors is on your right side at 1.5 meter". They should try to remember the location of the bottle. According to the description provided by the system, the user had to stand up and point towards the bottle.

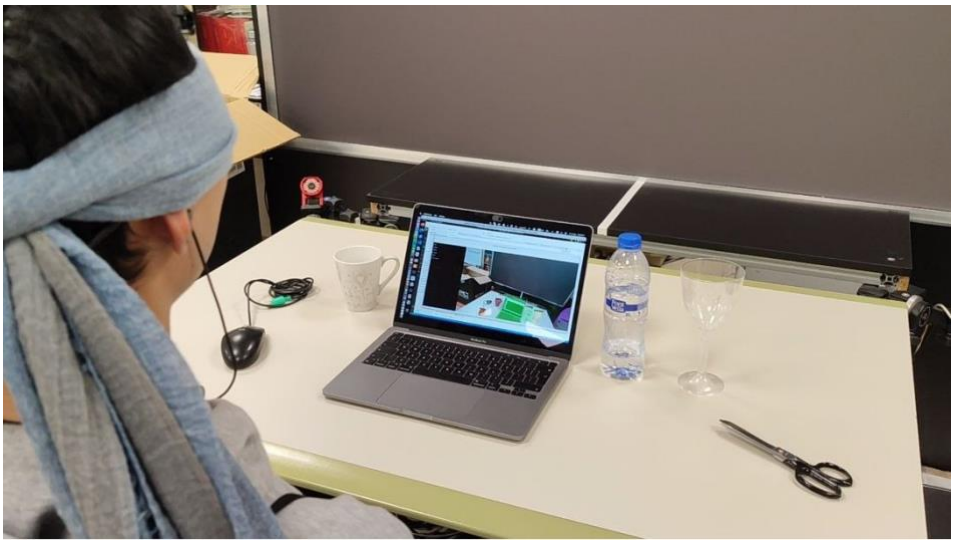

Figure 28 - Participant while doing the short distance object location testing scenario

8.1.1.4.2 Results

Three out of six participants in the test could not point correctly towards the position of the bottle. They pointed towards the front (at the laptop) instead of slightly to the right, where the bottle was located. However, the rest of the participants pointed approximately towards the intended object (bottle) (Figure 29).

Figure 29 - A participant while pointing at the intended object

8.1.2 Questionnaires

Once the users finished the testing process, they were asked to see the testing scenes without their eyes covered. They were asked to fill questionnaires containing some Likert scale questions and some open questions to answer. They had to rate their agreement with the statements on a scale from 1 to 5, where 1 was "Strongly Disagree" and 5 was "Strongly Agree."

The questions were as follows:

Scene understanding scenario

1. The system accurately detected and identified objects in the room.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
2. The voice user interface provided clear and useful information about the objects.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
3. The system's description of the scene was easy to understand.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
4. I felt confident in my ability to describe the scene based on the information provided by the system.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
5. The scene understanding mode was easy to use.
    1. Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree

Short distance object location scenario

1. The system accurately detected and identified objects in the short-distance scenario.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
2. The voice user interface provided clear and useful information about the objects' locations.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
3. I felt confident using the system in the short-distance scenario.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
4. The system was easy to use and interact with in the short-distance scenario.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree

Long distance object location scenario

1. The system accurately detected and identified objects in the long-distance scenario.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
2. The voice user interface provided clear and useful information about the objects' locations.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
3. The beeping mode effectively helped me locate objects in the long-distance scenario.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
4. I felt confident using the system in the long-distance scenario.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree
5. The system was easy to use and understand in the long-distance scenario.
   (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree

The open questions asked from the participants were as follows:

1. What aspects of the solution did you find most helpful in both short and long-distance scenarios?
2. What challenges or difficulties did you experience while using the solution in either scenario?
3. Do you have any suggestions for improvements or additional features that could enhance the solution's usability and effectiveness?
4. Please share any additional comments or feedback about your experience using the solution.

8.1.2.1 Questionnaires' results

The results of the questionnaires showed a high satisfaction of the user in various testing scenarios. The average satisfaction of the testers was 4 out of 5 in all testing scenarios. The answers of the participants are gathered in Table 18 - Questionnaires' results.

		P#1	P#2	P#3	P#4	P5	P6	Median
Scene understanding scenario	***Q1***	5	3	4	5	4	4	*4*
	Q2	4	4	3	4	5	5	*4*
	Q3	4	3	5	5	5	4	*4.5*
	Q4	4	4	3	4	4	3	*4*
	Q5	3	4	5	4	5	3	*4*
Short distance object location scenario	***Q1***	5	4	3	5	5	4	*4.5*
	Q2	4	3	4	5	5	5	*4.5*
	Q3	5	3	4	5	5	3	*4.5*
	Q4	3	3	5	4	5	3	*3.5*
Long distance object location scenario	***Q1***	3	4	4	5	3	4	*4*
	Q2	2	4	4	5	4	4	*4*
	Q3	2	2	5	5	5	4	*4.5*
	Q4	2	2	5	5	4	3	*3.5*
	Q5	3	2	5	5	4	4	*4*

Table 18 - Questionnaires' results

Participants also answered the open questions which are gathered in Table 19. Regarding the first question which was about the most helpful aspect of the solution, the description of the distance of the objects from them and the division of the scene into 3 parts (left, center and right) were the most helpful aspects for the testers since it helped them to build a better mental map of their surroundings. The second question was about the challenges that the participants had with the system demonstrated that the voice user interface interaction caused some difficulties for Participant 1 and 4. This is because participants had to wait for the system to indicate that it was listening before interacting with it. In some cases, users spoke the commands before the system indicated the listening mode, making the interaction problematic. P2 and P6 mentioned that they could not exactly understand where was the position of the objects and it would be better if they could get a more precise location of the objects instead of an approximate location. Regarding the improvements of the system in question three, P1 mentioned that specifying at which height the object is (knee level, chest level or head level) can help to understand better the 3D location of object in the scene. Moreover, other participants noted that more detailed information regarding the location of objects such as far right, far left would help to distinguish better the location of objects on each side. Lastly, some of their other comments concerned the system's response time, which could be improved by

using a faster or more precise beeping frequency to better understand distances between objects.

	P#1	P#2	P#3	P#4	P#5	P#6
Q1	It is really helpful to know if things were on the left, right and the distance	If they were on my left or right side	The beep is very good and the scene information is complete. I find it more useful for searching than for scanning.	The division between left, center and right	The description and distance of the objects. I also find it very useful to say if they are to the left, center or right.	Description of distance
Q2	Waiting to give a command and the speed of the beeping mode	When it says the objects is on your left/right I did not know how far on the left/right the objects were	It was difficult for me to imagine an entire scene instead of individual objects	I didn't know I should wait until he told me "listening" once I knew that everything was fine.	The greatest difficulty occurred in the long-distance scenario. because it took a long time to get a response	I don't know the detailed angel of where something is
Q3	Specify at which height the object is (knee level, chest level or head level)	Different terms for far-right, slight-left... etc.	Being able to choose which objects to say the distances for, otherwise with many objects it is too much	Saying the meters at which an object is helped me know which one was further away than the others, but I wouldn't know how to calculate without seeing the exact place where it was.	A description of the size and color of the object might be interesting. It would be useful if the interaction were faster, it could be made to recognize commands before the word "listening"	Direction can be described more detailed
Q4	It has been useful to locate the objects in the scene a little bit slow but worked correctly	It would be easier for the subject to give more context on the different tests he/she is going to do	Try to give "relative" positions (on the left a cup, on the right a computer...) to facilitate the creation of scenes			Maybe add something like scan button to beep at the same time can make it easier to understand scenarios also the beep frequency can be adjusted

Table 19 - Answers to the open questions

8.1.3 V-ASSISTANT preliminary user feedback

Since we had an agile approach in the development of the V-ASSISTANT we did not have the chance to perform an official testing like we did for AssistDiv. This was due to the difficulty of accessing the intended users and the short time we had for the development of the solution. However, up to the point of writing this thesis, we were able to access three persons with visual impairment or blindness and obtain some initial feedback from them after using the app.

One participant, provided by the Madrid Innovation Lab (Figure 30), had partial vision and relied on Android's accessibility features like TalkBack and high-contrast text. During the testing process, the focus was on the app's capability to describe scenes and assisting users in locating objects. The participant was particularly impressed by the system's use of an online API powered by generative AI, remarking that it "provides a more thorough explanation than a real person describing the scene.".

For locating objects, the app successfully helped the user identify and locate objects using a beeping mode, though she noted that the beeping mode has some delay, and it is a little bit confusing, especially when further from the object. Despite these minor issues, the participant expressed overall satisfaction with the app. Additionally, we had the chance to do preliminary testing at South Summit 2024 in Madrid provided further insights from two more testers. They emphasized the importance of having an offline mode for privacy and usability in areas with poor network coverage.

Figure 30 - A blind person testing V-ASSISTANT

The users also noted practical challenges, such as handling the phone while navigating with a cane or bag, and potential difficulties with TalkBack navigation. They suggested that a glasses-based interface might be more user-friendly.

It is important to mention that the process of the evaluation of the V-ASSITANT is still and ongoing process and we are looking forward to gathering more feedback from a variety of users in the near future.

9 Conclusion

This thesis was an effort to provide help for the researchers and developers in the field of assistive technologies for P-VI/blindness to design and develop more useful cost-efficient computer-vision based solutions for indoor environments. To achieve this goal, first we analyzed the state of the art through an SMS which highlighted the need for researchers and designers to focus more on the practical needs of people with visual impairments and the real-world applicability of assistive solutions. Many reviewed papers lacked context-specific data, making it difficult to assess the alignment of solutions with ICF categories and related daily tasks. While some solutions could detect objects or obstacles, their actual benefits for end users remained unclear, raising concerns about generalizability and the challenges of adapting solutions to different contexts. Embracing user-centered design (UCD) and the ICF framework, can lead to better-designed, more accepted solutions. Finally, the wide array of available software and hardware options complicates developers' decisions, stressing the need for a more structured approach in technology selection to build effective assistive solutions. As a result of this analysis, we identified some of the existing challenges in the development of these solutions, including:

- Integrating user needs and requirements into the design and development process
- Selecting appropriate technologies—both hardware and software—from a wide range of available options
- Effectively communicating feedback from the system to users.

Unlike many existing solutions, this thesis emphasizes integrating the user needs into the design process. This approach contrasts with other systems that might prioritize technological innovation over user experience. First we carried out a state-of-the-art analysis and then a series of semi-structured interviews with 8 participants that have different levels of visual impairment and blindness to better understand the user needs. The user research provided valuable insights into the needs and challenges faced by P-VI/blind individuals in indoor environments, revealing their reliance on a mix of traditional tools (e.g., white canes, guide dogs) and digital assistive technologies (e.g., Seeing AI, Be My Eyes, Google Maps). However, these solutions often lack accuracy, fail to address specific scene understanding needs, and raise privacy concerns. Key findings highlighted the need for better obstacle detection (including moving and upper-body obstacles), hierarchical information delivery to avoid cognitive overload, and

non-intrusive feedback modalities such as sound or vibration, with bone-conduction headphones particularly favoured for maintaining auditory awareness. Frequent challenges included locating misplaced objects, navigating unfamiliar spaces, and performing complex tasks like cooking or finding specific items in shared environments.

Afterwards, a framework was proposed to help solution designers, including several key components:

- The various use cases (defined based on the interviews and state-of-the-art analysis) of an assistive solution for scene understanding in indoor environments
- The list of functional and non-functional requirements to be fulfilled by such a solution
- A general reference architecture for assistive solutions for scene understanding in indoor environments
- A guideline for selecting appropriate technologies for the design and development of these solutions.

According to the framework, we evaluated an existing commercial solution (Seeing AI) as an example of how the framework can be used not only for design but also for the characterization of an already existing solution, or even for the comparison among several solutions. The Seeing AI app integrates several functional features, such as voice-based output, scene description, text detection, and specialized tools like barcode and currency readers. However, some functionalities considered in our use cases, such as object location and obstacle detection, were missing, possibly due to smartphone hardware limitations. While the app excels in text detection and provided innovative light and color detection features, areas like scene descriptions and face detection required further refinement. Non-functional evaluation highlighted the app's affordability, portability, and energy efficiency due to its smartphone-based design, though latency in scene descriptions and potential security risks from data sharing with Microsoft servers raise privacy concerns.

Additionally, two solutions were developed as secondary objectives of this thesis. AssistDiv was proposed to evaluate the potential existing technologies. It was developed on a laptop utilizing an RGB-D camera to get the environment information as input and provide scene understanding and object location assistance to the intended users. V-ASSISTANT was developed on a smartphone as an effort to develop a commercialized version of the solution. This was planned to fulfil the industrial side of the thesis. The solutions developed consider

user privacy by incorporating offline modes (in the case of V-ASSISTANT besides the generative AI API). This design choice ensures that sensitive data is processed locally on the device rather than being transmitted over the internet, thereby enhancing user privacy. Offline modes are particularly beneficial in assistive technologies as they not only protect user data but also improve accessibility by allowing the solutions to function without requiring a constant internet connection. This matter is overlooked in different existing solutions including BeMyEyes, Seeing AI and Envision. Additionally, our solutions utilize natural language for object location and scene understanding, in addition to a beeping mode. The use of natural language provides detailed verbal descriptions that convey object locations, allowing users to develop a cognitive map of their environment. This method contrasts with some state-of-the-art solutions [35], [82], [111] which often rely solely on auditory signals like beeps without providing verbal feedback. Moreover, V-ASSISTANT provides three different scene description modes (short/ long descriptions, and ask question) using the generative AI API which gives the user the possibility of choosing the desired option when necessary. Such flexibility does not exist in Seeing AI and Look Out. By combining the distance detection and two different scene description modes V-ASSISTANT surpasses its competitors, as outlined in Table 15.

It is important to mention that affordability was another important aspect that we had in mind while developing V-ASSISTANT. It is designed to operate on widely available Android devices, leveraging existing hardware without requiring additional expensive equipment. This enables a large number of users to have access to the solution, particularly those in low-income environments or areas with restricted access to subsidies for assistive technology.

Both solutions were tested and evaluated. The first solution (AssistDiv) was tested with blindfolded participants through an official testing. The second solution (V-ASSISTANT) underwent preliminary testing with P-VI/blindness. Across both evaluations, participants provided valuable insights and feedback, with an overall positive response to the functionality and usability of the solutions. The scene description feature in both solutions caught the attention of the users. Especially, the generative AI API used for scene description in V-ASSISTANT was surprising for them because of the detail it could provide about the scene. They also found object location useful for finding a specific object utilizing the beeps. However, it was expected to work more real-time in both solutions.

Our framework has demonstrated to be useful both for evaluation and analysis of current solutions and for the design of new ones. This work can guide and widen the horizons of researchers and developers who would like to contribute a scene understanding assistance solution. We reviewed different approaches that could be utilized to design such systems, as well as the challenges that could arise throughout the design process, and how an optimal solution should deal with the tradeoff between user requirements, available resources, and limitations of technologies.

10 Future work

There are various aspects of this thesis that could be investigated in the future. We have investigated the potential of assistive solutions for scene understanding by testing our prototypes with both blindfolded and P-VI/blindness participants. However, our solutions had limitations that affected the process of testing. For instance, the accuracy and response time of object detection models was one of the issues that affected the performance of the system (delay in beeping mode and the detection of objects in the scene). A more affective testing could be done by designing a wizard of Oz solution that mocks the system's behavior, allowing researchers to focus on understanding the user experience without being constrained by the technical limitations of the prototype. This approach would enable a more detailed investigation of how users interact with the assistive technology, providing insights that could guide future developments.

A more in-depth user evaluation is also essential to fully understand the impact and usability of the proposed assistive solutions. While tests with blindfolded and preliminary feedback from P-VI/blindness participants have provided valuable insights, these evaluations were limited in scope. Future studies should incorporate a broader range of users and explore long-term interactions with the system. This would offer a more comprehensive view of how well the solutions supports users' needs and reveal potential areas for improvement in terms of accessibility, ease of use, and overall user satisfaction.

Additionally, the way in which a solution provides feedback is a topic that deserves additional research out of the scope of this paper. For example, how current Multimodal LLMs provide descriptions, how they could be tailored for the P-VI/blindness, and how to deal with the hallucination issues of these models, are all important issues that must be addressed. More complex descriptions, such as describing objects in different colors, the semantic relationships between different objects, and how to convey all this information to the user via various modalities, should also be investigated in the future.

Furthermore, the privacy concerns that such solutions can raise in both users and those exposed to these technologies are worth considering. For example, a face detection feature could be integrated into the object detection module, but the entire process of obtaining and maintaining such data raises some concerns about data privacy that should be further investigated. It is also important to take into account the framework's compliance with user satisfaction in greater

detail, as we have discussed, to determine how compliance with the framework could impact the user experience. Additionally, this framework needs to be utilized and evaluated by developers and researchers to be refined and elaborated. This will provide valuable insights into areas for improvement of the framework to better serve its intended users.

11 References

[1] D. Pascolini and S. P. Mariotti, “Global estimates of visual impairment: 2010,” *British Journal of Ophthalmology*, vol. 96, no. 5, pp. 614 LP – 618, May 2012, doi: 10.1136/bjophthalmol-2011-300539.

[2] B. Kuipers, “The Cognitive Map: Could It Have Been Any Other Way?,” *Spatial Orientation*, no. September 1996, 1983, doi: 10.1007/978-1-4615-9325-6.

[3] S. Tan, D. Guo, H. Liu, X. Zhang, and F. Sun, “Embodied scene description,” *Auton Robots*, vol. 46, no. 1, pp. 21–43, 2022, doi: 10.1007/s10514-021-10014-9.

[4] K. Delloul and S. Larabi, “Egocentric Scene Description for the Blind and Visually Impaired,” in *2022 5th International Symposium on Informatics and its Applications (ISIA)*, 2022, pp. 1–6. doi: 10.1109/ISIA55826.2022.9993531.

[5] M. L. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan, “Fast indoor scene description for blind people with multiresolution random projections,” *J Vis Commun Image Represent*, vol. 44, pp. 95–105, 2017, doi: https://doi.org/10.1016/j.jvcir.2017.01.025.

[6] M. Hersh, “Mental Maps and the Use of Sensory Information by Blind and Partially Sighted People,” *ACM Trans Access Comput*, vol. 13, no. 2, 2020, doi: 10.1145/3375279.

[7] A. Belz, A. Muscat, M. Aberton, and S. Benjelloun, “Describing Spatial Relationships between Objects in Images in English and French,” Nov. 2015, pp. 104–113. doi: 10.18653/v1/W15-2816.

[8] A. Mishra and M. Liwicki, “Using Deep Object Features for Image Descriptions,” Nov. 2019. doi: 10.48550/arXiv.1902.09969.

[9] M. Nguyen, H. Le, W. Q. Yan, and A. Dawda, “A Vision Aid for the Visually Impaired using Commodity Dual-Rear-Camera Smartphones,” *Proceedings of the 2018 25th International Conference on Mechatronics and Machine Vision in Practice, M2VIP 2018*, vol. 1, pp. 8–13, 2019, doi: 10.1109/M2VIP.2018.8600857.

[10] W. Jeamwatthanachai, M. Wald, and G. Wills, “Indoor navigation by blind people: Behaviors and challenges in unfamiliar spaces and buildings,” *British Journal of Visual Impairment*, vol. 37, 2018, doi: 10.1177/0264619619833723.

[11] A. Alamri, “Development of Ontology-Based Indoor Navigation Algorithm for Indoor Obstacle Identification for the Visually Impaired,” *2023 9th International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, pp. 38–42, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:259266320

[12] S. Szpiro, Y. Zhao, and S. Azenkot, “Finding a Store, Searching for a Product: A Study of Daily Challenges of Low Vision People,” in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, in UbiComp ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 61–72. doi: 10.1145/2971648.2971723.

[13] M. M. Valipoor and A. de Antonio, “Recent trends in computer vision-driven scene understanding for VI/blind users: a systematic mapping,” *Univers Access Inf Soc*, Feb. 2022, doi: 10.1007/s10209-022-00868-w.

[14] K. Petersen, S. Vakkalanka, and L. Kuzniarz, “Guidelines for conducting systematic mapping studies in software engineering: An update,” *Inf Softw Technol*, vol. 64, pp. 1–18, 2015, doi: 10.1016/j.infsof.2015.03.007.

[15] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, “Systematic Mapping Studies in Software Engineering,” in *International Conference on Evaluation \& Assessment in Software Engineering*, 2008.

[16] E. Pell, Jr. L. E. Arend, and G. T. Timberlake, “Computerized Image Enhancement for Visually Impaired Persons: New Technology, New Possibilities,” *J Vis Impair Blind*, vol. 80, no. 7, pp. 849–854, 1986, doi: 10.1177/0145482X8608000709.

[17] E. Peli and T. Peli, “Image Enhancement For The Visually Impaired,” *Optical Engineering*, vol. 23, no. 1, pp. 2337–2350, 1984, doi: 10.1117/12.7973251.

[18] R. Sharma, M. Saqib, C. T. Lin, and M. Blumenstein, “A Survey on Object Instance Segmentation,” *SN Comput Sci*, vol. 3, no. 6, p. 499, 2022, doi: 10.1007/s42979-022-01407-3.

[19] C. Charisis and D. Argyropoulos, “Deep learning-based instance segmentation architectures in agriculture: A review of the scopes and challenges,” *Smart Agricultural Technology*, vol. 8, p. 100448, 2024, doi: https://doi.org/10.1016/j.atech.2024.100448.

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[21] C. Szegedy *et al.*, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[23] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, “Object Detection with Deep Learning: A Review,” *IEEE Trans Neural Netw Learn Syst*, vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865.

[24] “Amazon Rekognition.” Accessed: Feb. 05, 2024. [Online]. Available: https://aws.amazon.com/rekognition/

[25] D. Bharatia, P. Ambawane, and P. Rane, “Smart Electronic Stick for Visually Impaired using Android Application and Google’s Cloud Vision,” pp. 1–6, 2020, doi: 10.1109/gcat47503.2019.8978303.

[26] B. Jiang, J. Yang, Z. Lv, and H. Song, “Wearable vision assistance system based on binocular sensors for visually impaired users,” *IEEE Internet Things J*, vol. 6, no. 2, pp. 1375–1383, 2019, doi: 10.1109/JIOT.2018.2842229.

[27] S. Dosi, S. Sambare, S. Singh, N. Lokhande, and B. Garware, “Android Application for Object Recognition Using Neural Networks for the Visually Impaired,” in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1–6. doi: 10.1109/ICCUBEA.2018.8697886.

[28] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” 2017. [Online]. Available: https://arxiv.org/abs/1704.04861

[29] D. and E. D. and S. C. and R. S. and F. C.-Y. and B. A. C. Liu Wei and Anguelov, “SSD: Single Shot MultiBox Detector,” in *Computer Vision – ECCV 2016*, J. and S.

N. and W. M. Leibe Bastian and Matas, Ed., Cham: Springer International Publishing, 2016, pp. 21–37.

[30] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–14, 2015.

[31] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal Visual Object Classes (VOC) Challenge,” *Int J Comput Vis*, vol. 88, no. 2, pp. 303–338, 2010, doi: 10.1007/s11263-009-0275-4.

[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[33] “Redmon J Darknet: Open Source Neural Networks in C.” [Online]. Available: https://pjreddie.com/darknet/

[34] S. Duman, A. Elewi, and Z. Yetgin, “Design and implementation of an embedded real-Time system for guiding visually impaired individuals,” *2019 International Conference on Artificial Intelligence and Data Processing Symposium, IDAP 2019*, 2019, doi: 10.1109/IDAP.2019.8875942.

[35] M. Eckert, M. Blex, and C. M. Friedrich, “Object detection featuring 3d audio localization for microsoft hololens a deep learning based sensor substitution approach for the blind,” *HEALTHINF 2018 - 11th International Conference on Health Informatics, Proceedings; Part of 11th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2018*, vol. 5, no. Biostec, pp. 555–561, 2018, doi: 10.5220/0006655605550561.

[36] B. Kommey, K. Herrman, and E. O. Addo, “A Smart Vision Based Navigation Aid for the Visually Impaired,” *Asian Journal of Research in Computer Science*, no. November, pp. 1–8, 2019, doi: 10.9734/ajrcos/2019/v4i330114.

[37] M. Martinez, A. Roitberg, D. Koester, R. Stiefelhagen, and B. Schauerte, “Using Technology Developed for Autonomous Cars to Help Navigate Blind People,” *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops,*

ICCVW 2017, vol. 2018-Janua, pp. 1424–1432w, 2017, doi: 10.1109/ICCVW.2017.169.

[38] J. Wang, K. Yang, W. Hu, and K. Wang, "An Environmental Perception and Navigational Assistance System for Visually Impaired Persons Based on Semantic Stixels and Sound Interaction," *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018*, pp. 1921–1926, 2019, doi: 10.1109/SMC.2018.00332.

[39] U. and P. D. Badino Hernán and Franke, "The Stixel World - A Compact Medium Level Representation of the 3D-World," in *Pattern Recognition*, G. and S. H. Denzler Joachim and Notni, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 51–60.

[40] S. Bhumbla, D. K. Gupta, and Nisha, "A Review: Object Detection Algorithms," in *ICSCCC 2023 - 3rd International Conference on Secure Cyber Computing and Communications*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 827–832. doi: 10.1109/ICSCCC58608.2023.10176865.

[41] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

[42] A. K. Venkataramanan, M. Facktor, P. Gupta, and A. C. Bovik, "Assessing the impact of image quality on object-detection algorithms," *Electronic Imaging*, vol. 34, no. 9, pp. 334–1–334–1, 2022, doi: 10.2352/EI.2022.34.9.IQSP-334.

[43] H. Guo, T. Lu, and Y. Wu, "Dynamic Low-Light Image Enhancement for Object Detection via End-to-End Training," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 5611–5618. doi: 10.1109/ICPR48806.2021.9412802.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

[45] H. and R. J.-R. and W. E. and F. Y. Li Xiang and Cui, "Cross-Safe: A Computer Vision-Based Approach to Make All Intersection-Related Pedestrian Signals

Accessible for the Visually Impaired," in *Advances in Computer Vision*, S. Arai Kohei and Kapoor, Ed., Cham: Springer International Publishing, 2020, pp. 132–146.

[46] "Keras." Accessed: Jul. 29, 2024. [Online]. Available: https://keras.io/

[47] "OpenCV", Accessed: Jul. 29, 2024. [Online]. Available: https://opencv.org/

[48] "Matlab. Computer vision system toolbox2", Accessed: Jul. 29, 2024. [Online]. Available: https://www.mathworks.com/products/matlab.html

[49] J. Sosa-García and F. Odone, "'Hands on' visual recognition for visually impaired users," *ACM Transactions on Accessible Computingm*, vol. 10, no. 3, pp. 1–30, 2017, doi: 10.1145/3060056.

[50] H. Zhang and C. Ye, "An Indoor Wayfinding System Based on Geometric Features Aided Graph SLAM for the Visually Impaired," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 9, pp. 1592–1604, 2017, doi: 10.1109/TNSRE.2017.2682265.

[51] A. Vasconcelos Canez, J. Sartori, R. Barwaldt, and R. Nagel Rodrigues, "Collision detection with monocular vision for assisting in mobility of visually impaired people," *Proceedings - 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019*, pp. 269–274, 2019, doi: 10.1109/BRACIS.2019.00055.

[52] H. Hakim and A. Fadhil, "Navigation system for visually impaired people based on RGB-D camera and ultrasonic sensor," *ACM International Conference Proceeding Series*, pp. 172–177, 2019, doi: 10.1145/3321289.3321303.

[53] A. Bharati Puja and Pramanik, "Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey," in *Computational Intelligence in Pattern Recognition*, J. and N. B. and P. S. K. and P. D. Das Asit Kumar and Nayak, Ed., Singapore: Springer Singapore, 2020, pp. 657–668.

[54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[55] L. A. N. D. L. C. A. N. D. Y. D. A. N. D. T. Z. Gaihua Wang AND Jinheng, "Instance segmentation convolutional neural network based on multi-scale attention

mechanism," *PLoS One*, vol. 17, no. 1, pp. 1–14, Aug. 2022, doi: 10.1371/journal.pone.0263134.

[56] F. A. Miles, "Binocular Vision and Stereopsis by Ian P. Howard and Brian J. Rogers, Oxford University Press, 1995. £90.00 (736 pages) ISBN 0 19 508476 4.," *Trends Neurosci*, vol. 19, pp. 407–408, 1996.

[57] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, 2001, pp. 131–140. doi: 10.1109/SMBV.2001.988771.

[58] "Cámara Intel® RealSense™ SR300." Accessed: Feb. 08, 2024. [Online]. Available: https://www.intel.la/content/www/xl/es/products/sku/92329/intel-realsense-camera-sr300/specifications.html

[59] G. Maculotti *et al.*, "A methodology for task-specific metrological characterization of low-cost 3D camera for face analysis," *Measurement (Lond)*, vol. 200, Aug. 2022, doi: 10.1016/j.measurement.2022.111643.

[60] "A Brief Analysis of the Principles of Depth Cameras: Structured Light, TOF, and Stereo Vision." Accessed: Feb. 25, 2024. [Online]. Available: https://wiki.dfrobot.com/brief_analysis_of_camera_principles

[61] "Azure Kinect depth camera." Accessed: Feb. 22, 2022. [Online]. Available: https://azure.microsoft.com/en-us/products/kinect-dk

[62] "Azure Kinect DK hardware specifications." Accessed: Jan. 21, 2024. [Online]. Available: https://learn.microsoft.com/en-us/azure/kinect-dk/hardware-specification

[63] "Intel RealSense LiDAR Camera L515." Accessed: Jan. 21, 2024. [Online]. Available: https://intelrealsense.com/lidar-camera-l515/

[64] F. Khan, S. Salahuddin, and H. Javidnia, "Deep Learning-Based Monocular Depth Estimation Methods—A State-of-the-Art Review," *Sensors*, vol. 20, no. 8, Apr. 2020, doi: 10.3390/s20082272.

[65] J. M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, "CAM-Convs: Camera-Aware Multi-Scale Convolutions for Single-View Depth," in *2019*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11818–11827. doi: 10.1109/CVPR.2019.01210.

[66] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper Depth Prediction with Fully Convolutional Residual Networks," in *2016 Fourth International Conference on 3D Vision (3DV)*, Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2016, pp. 239–248. doi: 10.1109/3DV.2016.32.

[67] R. Wang, S. M. Pizer, and J. Frahm, "Recurrent Neural Network for (Un-)Supervised Learning of Monocular Video Visual Odometry and Depth," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2019, pp. 5550–5559. doi: 10.1109/CVPR.2019.00570.

[68] Y.-Z. Hsieh, S.-S. Lin, and F.-X. Xu, "Development of a wearable guide device based on convolutional neural network for blind or visually impaired persons," *Multimed Tools Appl*, vol. 79, no. 39, pp. 29473–29491, 2020, doi: 10.1007/s11042-020-09464-7.

[69] F. Khan, S. Salahuddin, and H. Javidnia, "Deep learning-based monocular depth estimation methods—a state-of-the-art review," *Sensors (Switzerland)*, vol. 20, no. 8, pp. 1–16, 2020, doi: 10.3390/s20082272.

[70] "Unsupervised Monocular Depth Estimation in Highly Complex Environments." Accessed: Nov. 29, 2022. [Online]. Available: https://en.x-mol.com/paper/article/1420832096734703616

[71] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, May 2021, doi: 10.1016/J.NEUCOM.2020.12.089.

[72] B. S. Lin, C. C. Lee, and P. Y. Chiang, "Simple smartphone-based guiding system for visually impaired people," *Sensors (Switzerland)*, vol. 17, no. 6, 2017, doi: 10.3390/s17061371.

[73] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Trans Pattern Anal Mach Intell*, vol. 29, no. 6, pp. 1052–1067, Jun. 2007, doi: 10.1109/TPAMI.2007.1049.

[74] W. M. Elmannai and K. M. Elleithy, “A novel obstacle avoidance system for guiding the visually impaired through the use of fuzzy control logic,” *CCNC 2018 - 2018 15th IEEE Annual Consumer Communications and Networking Conference*, vol. 2018-Janua, pp. 1–9, 2018, doi: 10.1109/CCNC.2018.8319310.

[75] P. Xu, G. A. Kennedy, F. Y. Zhao, W. J. Zhang, and R. Van Schyndel, “Wearable Obstacle Avoidance Electronic Travel Aids for Blind and Visually Impaired Individuals: A Systematic Review,” *IEEE Access*, vol. 11, no. June, pp. 66587–66613, 2023, doi: 10.1109/ACCESS.2023.3285396.

[76] A. F, A. NADA, M. A, and S. MASHALI, “Effective Fast Response Smart Stick for Blind People,” Institute of Research Engineers and Doctors, LLC, Apr. 2015, pp. 5–11. doi: 10.15224/978-1-63248-043-9-29.

[77] S. Khan, S. Nazir, and H. U. Khan, “Analysis of Navigation Assistants for Blind and Visually Impaired People: A Systematic Review,” *IEEE Access*, vol. 9, pp. 26712–26734, 2021, doi: 10.1109/ACCESS.2021.3052415.

[78] S. G. Jin, M. U. Ahmed, J. W. Kim, Y. H. Kim, and P. K. Rhee, “Combining obstacle avoidance and visual simultaneous localization and mapping for indoor navigation,” *Symmetry (Basel)*, vol. 12, no. 1, pp. 1–13, 2020, doi: 10.3390/SYM12010119.

[79] Z. Chen, X. Liu, M. Kojima, Q. Huang, and T. Arai, “A wearable navigation device for visually impaired people based on the real-time semantic visual slam system,” *Sensors*, vol. 21, no. 4, pp. 1–14, 2021, doi: 10.3390/s21041536.

[80] Y. Jia, X. Yan, and Y. Xu, “A Survey of simultaneous localization and mapping for robot,” in *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2019, pp. 857–861. doi: 10.1109/IAEAC47372.2019.8997820.

[81] O. Atoui, H. Husni, and R. C. Mat, “Visual-based semantic simultaneous localization and mapping for Robotic applications: A review,” *AIP Conf Proc*, vol. 2138, no. August, 2019, doi: 10.1063/1.5121082.

[82] C. Rui, Y. Liu, J. Shen, Z. Li, and Z. Xie, “A Multi-Sensory Blind Guidance System Based on YOLO and ORB-SLAM,” *Proceedings of the 2021 IEEE International*

Conference on Progress in Informatics and Computing, PIC 2021, pp. 409–414, 2021, doi: 10.1109/PIC53636.2021.9687018.

[83] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial, and Multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021, doi: 10.1109/TRO.2021.3075644.

[84] B. Kuriakose, R. Shrestha, and F. E. Sandnes, "SceneRecog: A Deep Learning Scene Recognition Model for Assisting Blind and Visually Impaired Navigate using Smartphones," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2021, pp. 2464–2470. doi: 10.1109/SMC52423.2021.9658913.

[85] P. G. Pawar and V. Devendran, "Scene understanding: A survey to see the world at a single glance," *2019 2nd International Conference on Intelligent Communication and Computational Techniques, ICCT 2019*, pp. 182–186, 2019, doi: 10.1109/ICCT46177.2019.8969051.

[86] L. Xie, F. Lee, L. Liu, K. Kotani, and Q. Chen, "Scene recognition: A comprehensive survey," *Pattern Recognit*, vol. 102, 2020, doi: 10.1016/j.patcog.2020.107205.

[87] N. Ketkar and J. Moolayil, "Convolutional Neural Networks," in *Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch*, N. Ketkar and J. Moolayil, Eds., Berkeley, CA: Apress, 2021, pp. 197–242. doi: 10.1007/978-1-4842-5364-9_6.

[88] Y. Liu, Q. Chen, W. Chen, and I. Wassell, "Dictionary Learning Inspired Deep Network for Scene Recognition," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, in AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[89] J. Shi, H. Zhu, S. Yu, W. Wu, and H. Shi, "Scene Categorization Model Using Deep Visually Sensitive Features," *IEEE Access*, vol. 7, pp. 45230–45239, 2019, doi: 10.1109/ACCESS.2019.2908448.

[90] H. Seong, J. Hyun, and E. Kim, “FOSNet: An end-to-end trainable deep neural network for scene recognition,” *IEEE Access*, vol. 8, pp. 82066–82077, 2020, doi: 10.1109/ACCESS.2020.2989863.

[91] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, 2006, pp. 2169–2178. doi: 10.1109/CVPR.2006.68.

[92] L.-J. Li and L. Fei-Fei, “What, where and who? Classifying events by scene and object recognition,” in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8. doi: 10.1109/ICCV.2007.4408872.

[93] A. Quattoni and A. Torralba, “Recognizing indoor scenes,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 413–420. doi: 10.1109/CVPR.2009.5206537.

[94] G. Patterson, C. Xu, H. Su, and J. Hays, “The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding,” *Int J Comput Vis*, vol. 108, no. 1, pp. 59–81, 2014, doi: 10.1007/s11263-013-0695-z.

[95] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., in Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 6105–6114.

[96] “Seeing AI.” Accessed: Dec. 14, 2023. [Online]. Available: https://www.microsoft.com/en-us/ai/seeing-ai

[97] Google, “Tesseract An optical character recognition (OCR) engine,” 2015. [Online]. Available: https://opensource.google/projects/tesseract

[98] N. Anwar, T. Khan, and A. F. Mollah, “Text Detection from Scene and Born Images: How Good is Tesseract?,” in *Recent Trends in Communication and Intelligent Systems*, A. K. S. Pundir, N. Yadav, H. Sharma, and S. Das, Eds., Singapore: Springer Nature Singapore, 2022, pp. 115–122.

[99] L. Neat, R. Peng, S. Qin, and R. Manduchi, “Scene Text Access: A Comparison of Mobile OCR Modalities for Blind Users,” in *Proceedings of the 24th International*

Conference on Intelligent User Interfaces, in IUI '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 197–207. doi: 10.1145/3301275.3302271.

[100] P. L. Kompalli, A. Kalidindi, J. Chilukala, K. Nerella, W. Shaik, and D. Cherukuri, "A Color Guide for Color Blind People Using Image Processing and OpenCV," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 09, pp. 30–46, Jul. 2023, doi: 10.3991/ijoe.v19i09.39177.

[101] I. and F. S. A. Allam Mahmoud and ElShaarawy, "Recognizing Clothing Patterns and Colors for BVI People Using Different Techniques," in *Digital Transformation Technology*, Y. K. and M. M. and J. A. Magdi Dalia A. and Helmy, Ed., Singapore: Springer Singapore, 2022, pp. 195–216.

[102] "Envision." Accessed: Dec. 14, 2023. [Online]. Available: https://www.letsenvision.com/

[103] "Lookout - Assisted vision." Accessed: Dec. 13, 2023. [Online]. Available: https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en&gl=US

[104] "Aira." Accessed: Dec. 14, 2023. [Online]. Available: https://aira.io/

[105] "Be My Eyes." Accessed: Dec. 14, 2023. [Online]. Available: https://www.bemyeyes.com/

[106] A. Stangl *et al.*, "Privacy Concerns for Visual Assistance Technologies," *ACM Trans. Access. Comput.*, vol. 15, no. 2, May 2022, doi: 10.1145/3517384.

[107] T. Norlund, L. Hagström, and R. Johansson, "Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?," *BlackboxNLP 2021 - Proceedings of the 4th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 149–162, 2021, doi: 10.18653/V1/2021.BLACKBOXNLP-1.10.

[108] C. Ntakolia, G. Dimas, and D. K. Iakovidis, "User-centered system design for assisted navigation of visually impaired individuals in outdoor cultural environments," *Univers Access Inf Soc*, no. 0123456789, 2020, doi: 10.1007/s10209-020-00764-1.

[109] H. Majerova, “The Aspects of Spatial Cognitive Mapping in Persons with Visual Impairment,” *Procedia Soc Behav Sci*, vol. 174, pp. 3278–3284, 2015, doi: 10.1016/j.sbspro.2015.01.994.

[110] A. M. Hafiz and G. M. Bhat, “A survey on instance segmentation: state of the art,” *Int J Multimed Inf Retr*, vol. 9, no. 3, pp. 171–189, 2020, doi: 10.1007/s13735-020-00195-x.

[111] G. Presti *et al.*, “Watchout: Obstacle sonifcation for people with visual impairment or blindness,” *ASSETS 2019 - 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 402–413, 2019, doi: 10.1145/3308561.3353779.

[112] Y. Liu, N. R. B. Stiles, and M. Meister, “Augmented reality powers a cognitive assistant for the blind,” *Elife*, vol. 7, Nov. 2018, doi: 10.7554/eLife.37841.

[113] A. Suresh, C. Arora, D. Laha, D. Gaba, and S. Bhambri, “Intelligent smart glass for visually impaired using deep learning machine vision techniques and robot operating system (ROS),” *Advances in Intelligent Systems and Computing*, vol. 751, pp. 99–112, 2019, doi: 10.1007/978-3-319-78452-6_10.

[114] S. Gandhi and N. Gandhi, “A CMUcam5 Computer Vision Based Arduino Wearable Navigation System for the Visually Impaired,” *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*, pp. 1768–1774, 2018, doi: 10.1109/ICACCI.2018.8554594.

[115] P. Vyavahare and S. Habeeb, “Assistant for Visually Impaired using Computer Vision,” *1st International Conference on Advanced Research in Engineering Sciences, ARES 2018*, pp. 1–7, 2018, doi: 10.1109/ARESX.2018.8723271.

[116] M. P. Arakeri, N. S. Keerthana, M. Madhura, A. Sankar, and T. Munnavar, “Assistive Technology for the Visually Impaired Using Computer Vision,” *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*, pp. 1725–1730, 2018, doi: 10.1109/ICACCI.2018.8554625.

[117] “Google. Google vision api.” Accessed: Aug. 08, 2024. [Online]. Available: https://cloud.google.com/vision

[118] S. Caraiman *et al.*, “Computer Vision for the Visually Impaired: The Sound of Vision System,” *Proceedings - 2017 IEEE International Conference on Computer Vision*

Workshops, ICCVW 2017, vol. 2018-Janua, pp. 1480–1489, 2017, doi: 10.1109/ICCVW.2017.175.

[119] "Google. Tesseract an optical character recognition (OCR) engine." Accessed: Aug. 08, 2024. [Online]. Available: https://opensource.google/projects/tesseract

[120] M. V Thomas and J. Abraham, "iSee : Artificial Intelligence Based Android Application for Visually Impaired People," vol. 21, no. 6, pp. 200–208, 2019.

[121] "Microsoft. Azure computer vision API", Accessed: Aug. 08, 2024. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

[122] F. Rahman, I. J. Ritun, N. Farhin, and J. Uddin, "AnAssistive model for visually impaired people using YOLO and MTCNN," *ACM International Conference Proceeding Series*, pp. 225–230, 2019, doi: 10.1145/3309074.3309114.

[123] "ICF. International classification of functioning, disability and health framework." Accessed: Aug. 09, 2024. [Online]. Available: https://apps.who.int/classifications/icfbrowser/

[124] J. H. Kim, S. K. Kim, T. M. Lee, Y. J. Lim, and J. Lim, "Smart glasses using deep learning and stereo camera," *2019 IEEE 8th Global Conference on Consumer Electronics, GCCE 2019*, vol. 2, pp. 294–295, 2019, doi: 10.1109/GCCE46687.2019.9015357.

[125] S. Pehlivan, M. Unay, and A. Akan, "Designing an obstacle detection and alerting system for visually impaired people on sidewalks," *TIPTEKNO 2019 - Tip Teknolojileri Kongresi*, pp. 1–4, 2019, doi: 10.1109/TIPTEKNO.2019.8895181.

[126] H. Alhichri, Y. Bazi, and N. Alajlan, "Assisting the Visually Impaired in Multi-object Scene Description Using OWA-Based Fusion of CNN Models," *Arab J Sci Eng*, vol. 45, no. 12, pp. 10511–10527, 2020, doi: 10.1007/s13369-020-04799-7.

[127] A. Aralikatti, J. Appalla, S. Kushal, G. S. Naveen, S. Lokesh, and B. S. Jayasri, "Real-time object detection and face recognition system to assist the visually impaired," *J Phys Conf Ser*, vol. 1706, no. 1, 2020, doi: 10.1088/1742-6596/1706/1/012149.

[128] S. Bhole and A. Dhok, "Deep Learning based Object Detection and Recognition Framework for the Visually-Impaired," in *Proceedings of the 4th International*

Conference on Computing Methodologies and Communication, ICCMC 2020, NIT, Dept. of ECE, Nagpur, India, 2020, pp. 725–728. doi: 10.1109/ICCMC48092.2020.ICCMC-000135.

[129] S. Malek, F. Melgani, M. L. Mekhalfi, and Y. Bazi, "Real-time indoor scene description for the visually impaired using autoencoder fusion strategies with visible cameras," *Sensors (Switzerland)*, vol. 17, no. 11, 2017, doi: 10.3390/s17112641.

[130] R. C. Joshi, S. Yadav, M. K. Dutta, and C. M. Travieso-Gonzalez, "Efficient Multi-Object Detection and Smart Navigation Using Artificial Intelligence for Visually Impaired People," *Entropy*, vol. 22, no. 9, 2020, doi: 10.3390/e22090941.

[131] B. Calabrese, R. Velázquez, C. Del-Valle-Soto, R. de Fazio, N. I. Giannoccaro, and P. Visconti, "Solar-powered deep learning-based recognition system of daily used objects and human faces for assistance of the visually impaired," *Energies (Basel)*, vol. 13, no. 22, 2020, doi: 10.3390/en13226104.

[132] H. C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarre, and D. Rus, "Enabling independent navigation for visually impaired people through a wearable vision-based feedback system," *Proc IEEE Int Conf Robot Autom*, pp. 6533–6540, 2017, doi: 10.1109/ICRA.2017.7989772.

[133] N. Mante and J. D. Weiland, "Visually Impaired Users can Locate and Grasp Objects under the Guidance of Computer Vision and Non-Visual Feedback," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2018-July, pp. 3493–3496, 2018, doi: 10.1109/EMBC.2018.8512918.

[134] J. Guerreiro, D. Sato, S. Asakawa, H. Dong, K. M. Kitani, and C. Asakawa, "Cabot: Designing and evaluating an autonomous navigation robot for blind people," *ASSETS 2019 - 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 68–82, 2019, doi: 10.1145/3308561.3353771.

[135] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Human Mental Workload*, vol. 52, P. A. Hancock and N. Meshkati, Eds., in Advances in Psychology, vol. 52. , North-Holland, 1988, pp. 139–183. doi: https://doi.org/10.1016/S0166-4115(08)62386-9.

[136] A. Budrionis, D. Plikynas, P. Daniušis, and A. Indrulionis, “Smartphone-based computer vision travelling aids for blind and visually impaired individuals: A systematic review,” *Assistive Technology*, vol. 0435, 2020, doi: 10.1080/10400435.2020.1743381.

[137] R. Verza, M. L. L. Carvalho, M. A. Battaglia, and M. M. Uccelli, “An interdisciplinary approach to evaluating the need for assistive technology reduces equipment abandonment,” *Multiple Sclerosis Journal*, vol. 12, no. 1, pp. 88–93, 2006.

[138] B. Phillips and H. Zhao, “Predictors of Assistive Technology Abandonment,” *Assistive Technology*, vol. 5, no. 1, pp. 36–45, 1993, doi: 10.1080/10400435.1993.10132205.

[139] H. Petrie, S. Carmien, and A. Lewis, *Assistive technology abandonment: Research realities and potentials*, vol. 10897 LNCS. Springer International Publishing, 2018. doi: 10.1007/978-3-319-94274-2_77.

[140] D. Townsend, F. Knoefel, and R. Goubran, “Privacy versus autonomy: A tradeoff model for smart home monitoring technologies,” in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 4749–4752. doi: 10.1109/IEMBS.2011.6091176.

[141] K. Lee, D. Sato, S. Asakawa, H. Kacorri, and C. Asakawa, “Pedestrian Detection with Wearable Cameras for the Blind: A Two-way Perspective,” *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–12, 2020, doi: 10.1145/3313831.3376398.

[142] T. Akter, “Privacy considerations of the visually impaired with camera based assistive tools,” *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, pp. 69–74, 2020, doi: 10.1145/3406865.3418382.

[143] “WeWalk. Smart cane for visually impaired and blind people.” Accessed: Aug. 11, 2024. [Online]. Available: https://wewalk.io/en/

[144] E. Wise *et al.*, “Indoor navigation for the blind and vision impaired: Where are we and where are we going?,” in *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, IEEE, Nov. 2012, pp. 1–7. doi: 10.1109/IPIN.2012.6418894.

[145] M. A. Hersh and M. A. Johnson, “A robotic guide for blind people. Part 1. A multi-national survey of the attitudes, requirements and preferences of potential end-users,” *Appl Bionics Biomech*, vol. 7, no. 4, pp. 277–288, 2010, doi: 10.1080/11762322.2010.523626.

[146] S. Ruffieux, C. Hwang, V. Junod, R. Caldara, D. Lalanne, and N. Ruffieux, “Tailoring assistive smart glasses according to pathologies of visually impaired individuals: an exploratory investigation on social needs and difficulties experienced by visually impaired individuals,” *Univers Access Inf Soc*, vol. 22, no. 2, pp. 463–475, Jun. 2023, doi: 10.1007/s10209-021-00857-5.

[147] P. Conradie, T. Mioch, and J. Saldien, “Blind user requirements to support tactile mobility,” *CEUR Workshop Proc*, vol. 1324, no. January 2015, 2014, doi: 10.13140/2.1.2492.3845.

[148] J. Madake, S. Bhatlawande, A. Solanke, and S. Shilaskar, “A Qualitative and Quantitative Analysis of Research in Mobility Technologies for Visually Impaired People,” *IEEE Access*, vol. 11, pp. 82496–82520, 2023, doi: 10.1109/ACCESS.2023.3291074.

[149] M. Mashiata *et al.*, “Towards assisting visually impaired individuals: A review on current status and future prospects,” Dec. 01, 2022, *Elsevier Ltd*. doi: 10.1016/j.biosx.2022.100265.

[150] W. Elmannai and K. Elleithy, “Sensor-Based Assistive Devices for Visually-Impaired People: Current Status, Challenges, and Future Directions,” *Sensors*, vol. 17, no. 3, 2017, doi: 10.3390/s17030565.

[151] J. Shen, Z. Dong, D. Qin, J. Lin, and Y. Li, *ivision: An assistive system for the blind based on augmented reality and machine learning*, vol. 12188 LNCS. Springer International Publishing, 2020. doi: 10.1007/978-3-030-49282-3_28.

[152] A. Khan and S. Khusro, *An insight into smartphone-based assistive solutions for visually impaired and blind people: issues, challenges and opportunities*, no. 0123456789. Springer Berlin Heidelberg, 2020. doi: 10.1007/s10209-020-00733-8.

[153] P. Conradie, T. Mioch, and J. Saldien, “Blind user requirements to support tactile mobility,” *CEUR Workshop Proc*, vol. 1324, no. January 2015, 2014, doi: 10.13140/2.1.2492.3845.

[154] K. Shinohara and J. O. Wobbrock, “In the Shadow of Misperception: Assistive Technology Use and Social Interactions,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, in CHI ’11. New York, NY, USA: Association for Computing Machinery, 2011, pp. 705–714. doi: 10.1145/1978942.1979044.

[155] A. D. P. dos Santos, A. L. M. Ferrari, F. O. Medola, and F. E. Sandnes, “Aesthetics and the perceived stigma of assistive technology for visual impairment,” *Disabil Rehabil Assist Technol*, vol. 17, no. 2, pp. 152–158, 2022, doi: 10.1080/17483107.2020.1768308.

[156] A. R. Hevner, “A three cycle view of design science research,” *Scandinavian journal of information systems*, vol. 19, no. 2, p. 4, 2007.

[157] V. Bajpai and R. P. Gorthi, “On Non-Functional Requirements : A Survey,” pp. 9–12, 2012.

[158] K. Tong and Y. Wu, “Deep learning-based detection from the perspective of small or tiny objects: A survey,” *Image Vis Comput*, vol. 123, p. 104471, 2022, doi: https://doi.org/10.1016/j.imavis.2022.104471.

[159] C. Silva and P. Wimalaratne, “Context-Aware Assistive Indoor Navigation of Visually Impaired Persons,” *Sensors and Materials*, vol. 32, p. 1497, Apr. 2020, doi: 10.18494/SAM.2020.2646.

[160] R. Tapu, B. Mocanu, and T. Zaharia, “Wearable assistive devices for visually impaired: A state of the art survey,” *Pattern Recognit Lett*, vol. 137, pp. 37–52, 2020, doi: 10.1016/j.patrec.2018.10.031.

[161] E. Ohn-Bar, K. Kitani, and C. Asakawa, “Personalized Dynamics Models for Adaptive Assistive Navigation Systems,” in *Conference on Robot Learning*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52937320

[162] L. Xie, F. Lee, L. Liu, Z. Yin, and Q. Chen, “Hierarchical Coding of Convolutional Features for Scene Recognition,” *IEEE Trans Multimedia*, vol. 22, no. 5, pp. 1182–1192, 2020, doi: 10.1109/TMM.2019.2942478.

[163] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors,” Institute of Electrical and Electronics Engineers (IEEE), Aug. 2023, pp. 7464–7475. doi: 10.1109/cvpr52729.2023.00721.

[164] “YoloV7-ncnn-Raspberry-Pi-4.” Accessed: Jan. 18, 2024. [Online]. Available: https://github.com/Qengineering/YoloV7-ncnn-Raspberry-Pi-4

[165] “Detectron2 Model Zoo and Baselines.” Accessed: Jan. 19, 2024. [Online]. Available: https://github.com/facebookresearch/detectron2/blob/main/MODEL_ZOO.md

[166] H. Caesar, J. Uijlings, and V. Ferrari, “COCO-Stuff: Thing and Stuff Classes in Context,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1209–1218. doi: 10.1109/CVPR.2018.00132.

[167] S. Cakic, T. Popovic, S. Krco, D. Nedic, D. Babic, and I. Jovovic, “Developing Edge AI Computer Vision for Smart Poultry Farms Using Deep Learning and HPC,” *Sensors*, vol. 23, no. 6, Mar. 2023, doi: 10.3390/s23063002.

[168] M. Cabanillas-Carbonell, A. A. Chávez, and J. B. Barrientos, “Glasses Connected to Google Vision that Inform Blind People about what is in Front of Them,” in *2020 International Conference on e-Health and Bioengineering (EHB)*, 2020, pp. 1–5. doi: 10.1109/EHB50910.2020.9280268.

[169] “Intel® RealSense™ Depth Camera D455.” Accessed: Jan. 19, 2024. [Online]. Available: https://www.intelrealsense.com/depth-camera-d455/

[170] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer,” *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 3, pp. 1623–1637, 2022, doi: 10.1109/TPAMI.2020.3019967.

[171] “MiDAS Github Repository.” Accessed: Jan. 21, 2024. [Online]. Available: https://github.com/isl-org/MiDaS

[172] M. Beshley, P. Volodymyr, H. Beshley, and M. Gregus, “A Smartphone-Based Computer Vision Assistance System with Neural Network Depth Estimation for the Visually Impaired,” 2023, pp. 26–36. doi: 10.1007/978-3-031-42508-0_3.

[173] R. and S. F. E. Kuriakose Bineeth and Shrestha, “LiDAR-Based Obstacle Detection and Distance Estimation in Navigation Assistance for Visually Impaired,” in *Universal Access in Human-Computer Interaction. User and Context Diversity*, C. Antona Margherita and Stephanidis, Ed., Cham: Springer International Publishing, 2022, pp. 479–491.

[174] S. Saranya, G. Sudha, and S. Subbiah, “Raspberry Pi based smart walking stick for visually impaired person,” *AIP Conf Proc*, vol. 2520, no. 1, p. 020010, Aug. 2022, doi: 10.1063/5.0103097.

[175] V. Kunta, C. Tuniki, and U. Sairam, “Multi-Functional Blind Stick for Visually Impaired People,” in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 895–899. doi: 10.1109/ICCES48766.2020.9137870.

[176] “Intel® RealSense™ Depth Camera D415.”

[177] S. S. Hussain, D. Durrani, A. A. Khan, R. Atta, and L. Ahmed, “In-door Obstacle Detection and Avoidance System for Visually Impaired People,” in *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, 2020, pp. 1–7. doi: 10.1109/GHTC46280.2020.9342942.

[178] M. Rajesh *et al.*, “Text recognition and face detection aid for visually impaired person using Raspberry PI,” in *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*, 2017, pp. 1–5. doi: 10.1109/ICCPCT.2017.8074355.

[179] C. Granquist, S. Y. Sun, S. R. Montezuma, T. M. Tran, R. Gage, and G. E. Legge, “Evaluation and Comparison of Artificial Intelligence Vision Aids: Orcam MyEye 1 and Seeing AI,” *J Vis Impair Blind*, vol. 115, no. 4, pp. 277–285, 2021, doi: 10.1177/0145482X211027492.

[180] J. Lee, J. Herskovitz, Y. H. Peng, and A. Guo, “ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image

Captions," in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, Apr. 2022. doi: 10.1145/3491102.3501966.

[181] G. Segalla, "Computer vision driven assistive solution for the visually impaired/blind people = Solución de asistencia impulsada por visión artificial para personas ciegas o con discapacidad visual," Jul. 2021. [Online]. Available: https://oa.upm.es/70185/

[182] M. and B. S. and H. J. and P. P. and R. D. and D. P. and Z. C. L. Lin Tsung-Yi and Maire, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, T. and S. B. and T. T. Fleet David and Pajdla, Ed., Cham: Springer International Publishing, 2014, pp. 740–755.

[183] F. and Z. J. and L. J. and Z. G. Chen Liang and Fang, "OID: Outlier Identifying and Discarding in Blind Image Deblurring," in *Computer Vision – ECCV 2020*, H. and B. T. and F. J.-M. Vedaldi Andrea and Bischof, Ed., Cham: Springer International Publishing, 2020, pp. 598–613.

[184] "Instance Segmentation With MiDaS Depth Detection." Accessed: Aug. 13, 2024. [Online]. Available: https://github.com/moev7/InstanceSegmentationWithDepth

[185] M. Servi *et al.*, "Metrological Characterization and Comparison of D415, D455, L515 RealSense Devices in the Close Range," *Sensors*, vol. 21, no. 22, 2021, doi: 10.3390/s21227770.

[186] C. Q. Zhao, Q. Y. Sun, C. Z. Zhang, Y. Tang, and F. Qian, "Monocular depth estimation based on deep learning: An overview," *Sci China Technol Sci*, vol. 63, no. 9, pp. 1612–1627, 2020, doi: 10.1007/s11431-020-1582-8.

[187] "AssistiDiv Github Repository." Accessed: Nov. 17, 2024. [Online]. Available: https://github.com/moev7/AssistDiv

[188] "SpeechRecognition." Accessed: Aug. 28, 2024. [Online]. Available: https://pypi.org/project/SpeechRecognition/

[189] "Vosk Speech Recognition Toolkit." Accessed: Aug. 28, 2024. [Online]. Available: https://github.com/alphacep/vosk-api

[190] "gTTS." Accessed: Aug. 28, 2024. [Online]. Available: https://gtts.readthedocs.io/en/latest/

[191] “Rapidfuzz.” Accessed: Aug. 28, 2024. [Online]. Available: https://github.com/rapidfuzz/RapidFuzz

[192] “Super Lidar.” Accessed: Oct. 11, 2024. [Online]. Available: https://www.supersense.app/product-page-superlidar

[193] “TalkBack.” Accessed: Aug. 28, 2024. [Online]. Available: https://support.google.com/accessibility/android/answer/6283677?hl=en

[194] “Anthropic API.” Accessed: Aug. 28, 2024. [Online]. Available: https://www.anthropic.com/api

[195] “AR Core Depth API.” [Online]. Available: https://developers.google.com/ar/develop/depth

[196] “EfficientDet.” Accessed: Oct. 07, 2024. [Online]. Available: https://github.com/google/automl/blob/master/efficientdet/README.md

[197] G. A. Miller, “The magical number seven, plus or minus two: Some limits on our capacity for processing information.,” *Psychol Rev*, vol. 63, no. 2, pp. 81–97, Mar. 1956, doi: 10.1037/h0043158.

[198] “Madrid HCI Lab UPM.” Accessed: Oct. 06, 2024. [Online]. Available: https://www.madhcilab.es/

[199] B. Mocanu, R. Tapu, and T. Zaharia, “Seeing Without Sight - An Automatic Cognition System Dedicated to Blind and Visually Impaired People,” *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, vol. 2018-Janua, pp. 1452–1459, 2017, doi: 10.1109/ICCVW.2017.172.

[200] Y. Zhao, R. Huang, and B. Hu, “A Multi-Sensor Fusion System for Improving Indoor Mobility of the Visually Impaired,” *Proceedings - 2019 Chinese Automation Congress, CAC 2019*, pp. 2950–2955, 2019, doi: 10.1109/CAC48633.2019.8996578.

[201] M. Iwamura, Y. Inoue, K. Minatani, and K. Kise, *Suitable camera and rotation navigation for people with visual impairment on looking for something using object detection technique*, vol. 12376 LNCS, no. 17. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58796-3_57.

[202] E. Yohannes, P. Lin, C. Y. Lin, and T. K. Shih, “Robot Eye: Automatic Object Detection and Recognition Using Deep Attention Network to Assist Blind People,” *Proceedings - 2020 International Conference on Pervasive Artificial Intelligence, ICPAI 2020*, pp. 152–157, 2020, doi: 10.1109/ICPAI51961.2020.00036.

[203] M. Afif, R. Ayachi, E. Pissaloux, Y. Said, and M. Atri, “Indoor objects detection and recognition for an ICT mobility assistance of visually impaired people,” *Multimed Tools Appl*, vol. 79, no. 41–42, pp. 31645–31662, 2020, doi: 10.1007/s11042-020-09662-3.

[204] V. N. Mandhala, D. Bhattacharyya, B. Vamsi, and N. Thirupathi Rao, “Object detection using machine learning for visually impaired people,” *Int J Curr Res Rev*, vol. 12, no. 20, pp. 157–167, 2020, doi: 10.31782/IJCRR.2020.122032.

[205] L. Abraham, N. S. Mathew, L. George, and S. S. Sajan, “VISION- Wearable Speech Based Feedback System for the Visually Impaired using Computer Vision,” in *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, Saintgits College of Engineering, Computer Science and Engineering Department, India, 2020, pp. 972–976. doi: 10.1109/ICOEI48184.2020.9142984.

[206] S. Vaidya, N. Shah, N. Shah, and R. Shankarmani, “Real-Time Object Detection for Visually Challenged People,” in *Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, Sardar Patel Institute of Technology, Information Technology Department, Mumbai, India, 2020, pp. 311–316. doi: 10.1109/ICICCS48265.2020.9121085.

[207] A. Kandoth, N. R. Arya, P. R. Mohan, T. V. Priya, and M. Geetha, “Dhrishti: A visual aiding system for outdoor environment,” *Proceedings of the 5th International Conference on Communication and Electronics Systems, ICCES 2020*, no. Icces, pp. 305–310, 2020, doi: 10.1109/ICCES48766.2020.09137967.

[208] M. Kuribayashi, S. Kayukawa, H. Takagi, C. Asakawa, and S. Morishima, “LineChaser: A Smartphone-Based Navigation System for Blind People to Stand in Lines,” pp. 1–13, 2021, doi: 10.1145/3411764.3445451.

[209] S. Suny, S. Basak, S. Mazharul, and S. M. M. H. Chowdhury, “Virtual Vision for Blind People Using Mobile Camera and Sonar sensors,” Jul. 2019.

[210] S. Kayukawa, T. Ishihara, H. Takagi, S. Morishima, and C. Asakawa, “BlindPilot: A robotic local navigation system that leads blind people to a landmark object,” *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–9, 2020, doi: 10.1145/3334480.3382925.

[211] L. Wang, A. Patnik, E. Wong, J. Wong, and A. Wong, “OLIV : An Artificial Intelligence-Powered Assistant for Object Localization for Impaired Vision,” *Journal of computational vision and imaging systems*, p. 3, 2018.

[212] S. Gianani, A. Mehta, T. Motwani, and R. Shende, “JUVO - An Aid for the Visually Impaired,” *2018 International Conference on Smart City and Emerging Technology, ICSCET 2018*, pp. 1–4, 2018, doi: 10.1109/ICSCET.2018.8537270.

[213] H. Nguyen, M. Nguyen, Q. Nguyen, S. Yang, and H. Le, “Web-based object detection and sound feedback system for visually impaired people,” *2020 International Conference on Multimedia Analysis and Pattern Recognition, MAPR 2020*, pp. 20–25, 2020, doi: 10.1109/MAPR49794.2020.9237770.

[214] Q. Chen, Y. Chen, J. Zhu, G. De Luca, M. Zhang, and Y. Guo, “Traffic light and moving object detection for a guide-dog robot,” *The Journal of Engineering*, vol. 2020, no. 13, pp. 675–678, 2020, doi: 10.1049/joe.2019.1137.

[215] D. P. Khairnar, R. B. Karad, A. Kapse, G. Kale, and P. Jadhav, “PARTHA: A Visually Impaired Assistance System,” in *2020 3rd International Conference on Communication Systems, Computing and IT Applications, CSCITA 2020 - Proceedings*, Pune Institute of Computer Technology, Pune, India, 2020, pp. 32–37. doi: 10.1109/CSCITA47329.2020.9137791.

[216] J. A. Shah, A. Raorane, A. Ramani, H. Rami, and N. Shekokar, “EYERIS: A Virtual Eye to Aid the Visually Impaired,” in *2020 3rd International Conference on Communication Systems, Computing and IT Applications, CSCITA 2020 - Proceedings*, D J Sanghvi College of Engineering, Departmente of Computer Engineering, Mumbai, India, 2020, pp. 202–207. doi: 10.1109/CSCITA47329.2020.9137777.

[217] R. Boldu, D. J. C. Matthies, H. Zhang, and S. Nanayakkara, “Aisee: An assistivewearable device to support visually impaired grocery shoppers,” *Proc ACM Interact Mob Wearable Ubiquitous Technol*, vol. 4, no. 4, 2020, doi: 10.1145/3432196.

[218] N. Tahoun, A. Awad, and T. Bonny, “Smart assistant for blind and visually impaired people,” *ACM International Conference Proceeding Series*, pp. 227–231, 2019, doi: 10.1145/3369114.3369139.

[219] P. Akkapusit and I. Y. Ko, “Task-oriented approach to guide visually impaired people during smart device usage,” *Proceedings - 2021 IEEE International Conference on Big Data and Smart Computing, BigComp 2021*, pp. 28–35, 2021, doi: 10.1109/BigComp51126.2021.00015.

[220] H. Baskaran, R. L. M. Leng, F. A. Rahim, and M. E. Rusli, “Smart Vision: Assistive Device for the Visually Impaired Community U sing Online Computer Vision Service,” in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 2019, pp. 730–734. doi: 10.1109/CCOMS.2019.8821635.

[221] M. Saha, A. J. Fiannaca, M. Kneisel, E. Cutrell, and M. R. Morris, “Closing the gap: Designing for the last-few-meters wayfinding problem for people with visual impairments,” *ASSETS 2019 - 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 222–235, 2019, doi: 10.1145/3308561.3353776.

[222] C. S. Silva and P. Wimalaratne, “Towards a grid based sensor fusion for visually impaired navigation using sonar and vision measurements,” *5th IEEE Region 10 Humanitarian Technology Conference 2017, R10-HTC 2017*, vol. 2018-Janua, pp. 784–787, 2018, doi: 10.1109/R10-HTC.2017.8289073.

[223] R. S. Dasila, M. Trivedi, S. Soni, M. Senthil, and M. Narendran, “Real time environment perception for visually impaired,” *Proceedings - 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development, TIAR 2017*, vol. 2018-Janua, pp. 168–172, 2018, doi: 10.1109/TIAR.2017.8273709.

[224] “Clarifi.” Accessed: Sep. 13, 2024. [Online]. Available: https://www.clarifai.com/

[225] “Cloud Sight.” Accessed: Sep. 13, 2024. [Online]. Available: https://cloudsight.ai/

[226] L. Tepelea, I. Buciu, C. Grava, I. Gavrilut, and A. Gacsadi, “A vision module for visually impaired people by using raspberry PI platform,” *2019 15th International Conference on Engineering of Modern Electric Systems, EMES 2019*, pp. 209–212, 2019, doi: 10.1109/EMES.2019.8795205.

[227] M. Afif, R. Ayachi, Y. Said, E. Pissaloux, and M. Atri, “An Evaluation of RetinaNet on Indoor Object Detection for Blind and Visually Impaired Persons Assistance Navigation,” *Neural Process Lett*, vol. 51, no. 3, pp. 2265–2279, 2020, doi: 10.1007/s11063-020-10197-9.

[228] A. Shelton and T. Ogunfunmi, “Developing a Deep Learning-enabled Guide for the Visually Impaired,” *2020 IEEE Global Humanitarian Technology Conference, GHTC 2020*, 2020, doi: 10.1109/GHTC46280.2020.9342873.

[229] “TensorFlow. Object detection.” Accessed: Aug. 21, 2024. [Online]. Available: https://www.tensorflow.org/

[230] S. Suresh and Akhilaa, “Vision: Android Application for the Visually Impaired,” *2020 IEEE International Conference for Innovation in Technology, INOCON 2020*, pp. 8–13, 2020, doi: 10.1109/INOCON50539.2020.9298325.

[231] M. and B. A. S. Islam Md. Tobibul and Ahmad, “Microprocessor-Based Smart Blind Glass System for Visually Impaired People,” in *Proceedings of International Joint Conference on Computational Intelligence*, J. C. Uddin Mohammad Shorif and Bansal, Ed., Singapore: Springer Nature Singapore, 2020, pp. 151–161.

[232] J. M. P. Barroso, “Safe and Sound Mobile Application,” pp. 22–28, 2020.

[233] M. A. Imtiaz, S. Aziz, A. Zaib, A. Maqsood, M. U. Khan, and A. Waseem, “Wearable Scene Classification System for Visually Impaired Individuals,” *2nd International Conference on Electrical, Communication and Computer Engineering, ICECCE 2020*, no. June, 2020, doi: 10.1109/ICECCE49384.2020.9179439.

[234] R. Cheng, K. Wang, J. Bai, and Z. Xu, “Unifying Visual Localization and Scene Recognition for People with Visual Impairment,” *IEEE Access*, vol. 8, pp. 64284–64296, 2020, doi: 10.1109/ACCESS.2020.2984718.

[235] F. and M. P. and C. E. and N. S. and K. I. Georgiadis Kostas and Kalaganis, “A Computer Vision System Supporting Blind People - The Supermarket Case,” in *Computer Vision Systems*, D. and V. M. and A. A. Tzovaras Dimitrios and Giakoumis, Ed., Cham: Springer International Publishing, 2019, pp. 305–315.

[236] B. R. and J. K. and K. M. S. and Y. V. Akula Rajani and Sai, “Efficient Obstacle Detection and Guidance System for the Blind (Haptic Shoe),” in *Advances in Decision*

Sciences, Image Processing, Security and Computer Vision, K. S. and S. K. and K. D. R. and F. M. N. Satapathy Suresh Chandra and Raju, Ed., Cham: Springer International Publishing, 2020, pp. 266–271.

[237] F. Breve and C. N. Fischer, "Visually Impaired Aid using Convolutional Neural Networks, Transfer Learning, and Particle Competition and Cooperation," *ArXiv*, 2020.

[238] Z. Chen, X. Liu, M. Kojima, Q. Huang, and T. Arai, "A wearable navigation device for visually impaired people based on the real-time semantic visual slam system," *Sensors*, vol. 21, no. 4, pp. 1–14, 2021, doi: 10.3390/s21041536.

[239] A. D. P. dos Santos, F. O. Medola, M. J. Cinelli, A. R. Garcia Ramirez, and F. E. Sandnes, "Are electronic white canes better than traditional canes? A comparative study with blind and blindfolded participants," *Univers Access Inf Soc*, vol. 20, no. 1, pp. 93–103, 2020, doi: 10.1007/s10209-020-00712-z.

[240] Y. Endo, K. Sato, A. Yamashita, and K. Matsubayashi, "Indoor positioning and obstacle detection for visually impaired navigation system based on LSD-SLAM," *Proceedings of 2017 International Conference on Biometrics and Kansei Engineering, ICBAKE 2017*, pp. 158–162, 2017, doi: 10.1109/ICBAKE.2017.8090635.

[241] J. Gay *et al.*, "Keep Your Distance: A Playful Haptic Navigation Wearable for Individuals with Deafblindness," *ASSETS 2020 - 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 8–10, 2020, doi: 10.1145/3373625.3418048.

[242] "Aruco: a minimal library for augmented reality applications based on OpenCV." Accessed: Aug. 12, 2024. [Online]. Available: https://www.uco.es/investiga/grupos/ava/node/26

[243] F. Huppert, G. Hoelzl, and M. Kranz, "GuideCopter - A Precise Drone-Based Haptic Guidance Interface for Blind or Visually Impaired People," pp. 1–14, 2021, doi: 10.1145/3411764.3445676.

[244] "Optitrack." Accessed: Aug. 21, 2024. [Online]. Available: https://optitrack.com/

[245] S. M. T. Islam, B. Woldegebriel, and A. Ashok, “TaxSeeMe: A Taxi Administering System for the Visually Impaired,” *IEEE Vehicular Networking Conference, VNC*, vol. 2018-Decem, pp. 1–2, 2019, doi: 10.1109/VNC.2018.8628328.

[246] S. Kayukawa, H. Takagi, J. Guerreiro, S. Morishima, and C. Asakawa, “Smartphone-based assistance for blind people to stand in lines,” *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–8, 2020, doi: 10.1145/3334480.3382954.

[247] S. Kayukawa, T. Ishihara, H. Takagi, S. Morishima, and C. Asakawa, “Guiding Blind Pedestrians in Public Spaces by Understanding Walking Behavior of Nearby Pedestrians,” *Proc ACM Interact Mob Wearable Ubiquitous Technol*, vol. 4, no. 3, 2020, doi: 10.1145/3411825.

[248] R. Tapu, B. Mocanu, and T. Zaharia, “A computer vision-based perception system for visually impaired,” *Multimed Tools Appl*, vol. 76, no. 9, pp. 11771–11807, 2017, doi: 10.1007/s11042-016-3617-6.

[249] A. K. Srinivasan, S. Sridharan, and R. Sridhar, “Object Localization and Navigation Assistant for the Visually challenged,” in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 324–328. doi: 10.1109/ICCMC48092.2020.ICCMC-00061.

[250] J. B. F. van Erp, L. C. M. Kroon, T. Mioch, and K. I. Paul, “Obstacle detection display for visually impaired: Coding of direction, distance, and height on a vibrotactile waist band,” *Frontiers in ICT*, vol. 4, no. SEP, pp. 1–19, 2017, doi: 10.3389/fict.2017.00023.

[251] S. Ur Rahman, S. Ullah, and S. Ullah, “A mobile camera based navigation system for visually impaired people,” *ACM International Conference Proceeding Series*, pp. 63–66, 2019, doi: 10.1145/3330180.3330193.

[252] B. Li *et al.*, “Vision-Based Mobile Indoor Assistive Navigation Aid for Blind People,” *IEEE Trans Mob Comput*, vol. 18, no. 3, pp. 702–714, 2019, doi: 10.1109/TMC.2018.2842751.

[253] S. and S. V. and S. S. Megalingam Rajesh Kannan and Vishnu, “Autonomous Path Guiding Robot for Visually Impaired People,” in *Cognitive Informatics and Soft Computing*, V. E. and B. A. K. and Z. A. F. Mallick Pradeep Kumar and Balas, Ed., Singapore: Springer Singapore, 2019, pp. 257–266.

[254] L. Wang, J. Zhao, and L. Zhang, “NavDog: Robotic navigation guide dog via model predictive control and human-robot modeling,” *Proceedings of the ACM Symposium on Applied Computing*, pp. 815–818, 2021, doi: 10.1145/3412841.3442098.

[255] B. Mocanu, R. Tapu, and T. Zaharia, “Seeing Without Sight — An Automatic Cognition System Dedicated to Blind and Visually Impaired People,” in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 1452–1459. doi: 10.1109/ICCVW.2017.172.

[256] J. R. Rizzo, Y. Pan, T. Hudson, E. K. Wong, and Y. Fang, “Sensor fusion for ecologically valid obstacle identification: Building a comprehensive assistive technology platform for the visually impaired,” *2017 7th International Conference on Modeling, Simulation, and Applied Optimization, ICMSAO 2017*, 2017, doi: 10.1109/ICMSAO.2017.7934891.

[257] R. Kedia, K. K. Yoosuf, P. Dedeepya, M. Fazal, C. Arora, and M. Balakrishnan, “MAVI: An Embedded Device to Assist Mobility of Visually Impaired,” *Proceedings - 2017 30th International Conference on VLSI Design and 2017 16th International Conference on Embedded Systems, VLSID 2017*, pp. 213–218, 2017, doi: 10.1109/VLSID.2017.38.

[258] H. Jabnoun, F. Benzarti, and H. Amiri, “Visual scene prediction for blind people based on object recognition,” *Proceedings - 2017 14th International Conference on Computer Graphics, Imaging and Visualization, CGiV 2017*, pp. 21–26, 2018, doi: 10.1109/CGiV.2017.19.

[259] M. Hudec and Z. Smutny, “Advanced scene recognition system for blind people in household the use of notification sounds in spatial and social context of blind people,” *ACM International Conference Proceeding Series*, pp. 1–5, 2018, doi: 10.1145/3207677.3278101.

[260] M. Awad, J. El Haddad, E. Khneisser, T. Mahmoud, E. Yaacoub, and M. Malli, “Intelligent eye: A mobile application for assisting blind people,” *2018 IEEE Middle East and North Africa Communications Conference, MENACOMM 2018*, pp. 1–6, 2018, doi: 10.1109/MENACOMM.2018.8371005.

[261] Y. Bazi, H. Alhichri, N. Alajlan, and F. Melgani, “Scene description for visually impaired people with multi-label convolutional svm networks,” *Applied Sciences (Switzerland)*, vol. 9, no. 23, 2019, doi: 10.3390/app9235062.

[262] A. Ghosh, S. A. Al Mahmud, T. I. R. Uday, and D. M. Farid, “Assistive Technology for Visually Impaired using Tensor Flow Object Detection in Raspberry Pi and Coral USB Accelerator,” *2020 IEEE Region 10 Symposium, TENSYMP 2020*, no. June, pp. 186–189, 2020, doi: 10.1109/TENSYMP50017.2020.9230630.

[263] G. Fusco and J. M. Coughlan, “Indoor localization for visually impaired travelers using computer vision on a smartphone,” *Proceedings of the 17th International Web for All Conference, W4A 2020*, 2020, doi: 10.1145/3371300.3383345.

[264] D. Ahmetovic, D. Sato, U. Oh, T. Ishihara, K. Kitani, and C. Asakawa, “ReCog: Supporting Blind People in Recognizing Personal Objects,” in *Conference on Human Factors in Computing Systems - Proceedings*, Università Degli Studi di Milano, Milano, Italy, 2020. doi: 10.1145/3313831.3376143.

[265] S. Ahmed, H. Balasubramanian, S. Stumpf, C. Morrison, A. Sellen, and M. Grayson, “Investigating the intelligibility of a computer vision system for blind users,” *International Conference on Intelligent User Interfaces, Proceedings IUI*, pp. 419–429, 2020, doi: 10.1145/3377325.3377508.

[266] M. G. Sarwar, A. Dey, and A. Das, “Developing a LBPH-based Face Recognition System for Visually Impaired People,” *2021 1st International Conference on Artificial Intelligence and Data Analytics, CAIDA 2021*, no. April, pp. 286–289, 2021, doi: 10.1109/CAIDA51941.2021.9425275.

[267] L. Stearns and A. Thieme, “Automated person detection in dynamic scenes to assist people with vision impairments: An initial investigation,” *ASSETS 2018 - Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 391–394, 2018, doi: 10.1145/3234695.3241017.

12 Appendix

Object detection method	Paper
YOLO Variations	Liu et al. 2018 [112], Rahman et al. 2019 [122], Kim et al. 2019 [124] Eckert et al. 2018 [35], Duman et al. 2019 [34] Kommey et al. 2019 [36], Lin et al. 2017 [72] Tapu et al. 2017 [199], Zhao et al. 2019 [200] Guerreiro et al. 2019 [134], Iwamura et al. 2020 [201] Yohannes et al. 2020 [202], Martinez et al. 2020 [203] Mandhala et al. 2020 [204], Abraham et al. 2020 [205], Aralikatti et al. 2020 [127], Vaidya et al. 2020 [206], Joshi et al. 2020 [130], Kandoth et al. 2020 [207], Vaidya et al. 2020 [206], Kuribayashi et al. 2021[208], Shen et al. 2020 [151], Suny et al. 2020 [209], Kayukawa et al. 2020 [210]
SSD Variations	Wang et al. 2018 [211], Suresh et al. 2019 [113], Pehlivan et al. 2019 [125], Gianani et al. 2018 [212], Hussain et al. 2020 [177], Nguyen et al. 2020 [213], Bhole and Dhok 2020 [128], Chen et al. 2020 [214], Khairnar et al. 2020 [215], Shah et al. 2020 [216], Boldu et al. 2020 [217], Tahoun et al. 2019 [218], Akkapusit and Ko 2021 [219]
Microsoft Azure Computer Vision	Baskaran et al. 2019 [220], Saha et al. 2019 [221], Thomas et al. [120], Vyava-hare and Habeeb 2018 [115]
Google Cloud Vision	Thomas et al. 2019 [120], Silva and Wimalaratne 2017 [222], Dasila et al. 2017 [223], Arakeri et al. 2018 [116], Bharatia et al. [25]
Clarifi [224]	Thomas et al. 2019 [120]
Cloud Sight API [225]	Thomas et al. 2019 [120]
SGD Algorithm on Keras	Li et al. 2019 [45]
OpenCV	Tepelea et al. 2019 [226]
Computer Vision System Tool-box2 from MATLAB	Sosa-García and Odone 2017 [49]
VGGNet	Afif et al. 2020 [227]
VGG16	Alhichri et al. 2020 [126]
AlexNet	Shelton and Ogunfunmi 2020 [228]
TensorFlow API [229]	Suresh et al. 2020 [230], Islam et al. 2020 [231], Eskicioglu et al. 2020 [232]
Quadratic Discriminant	Imtiaz et al. 2020 [233]

MobileNet v2	Cheng et al. 2020 [234]
ResNet	Georgiadis et al. 2019 [235]

Appendix 1 - Object detection methods

Article	Distance sensor-based	Camera-based	Detection approach
Akula et al. [236]	✓		Ultrasonic sensor
Arakeri et al. [116]	✓		Ultrasonic sensor
Bharatia et al. [25]	✓		Ultrasonic sensor
Breve and Fischer [237]		✓	A deep learning model for detecting obstacles
Canez et al. [51]		✓	Time To Collision (TTC) technique
Caraiman et al. [118]	✓	✓	3D reconstruction of the sensed environment using IR-based depth sensor and a stereo vision system
Chen et al. [238]		✓	RGB-D camera
Cheng et al. [234]		✓	Intel RealSense D435 depth camera
Dasila et al. [223]		✓	Cloud points
dos Santos et al. [239]	✓		Ultrasonic sensor
Elmannai and Elleithy [74]		✓	Using fuzzy logic
Endo et al. [240]		✓	SLAM
Gandhi and Gandhi [114]		✓	Pixy Cam
Gay et al. [241]		✓	ArUco Marker [242] and OpenCV [47]
Guerreiro et al. [134]	✓	✓	LiDAR sensor and ZED RGB-D camera
Hakim and Fadhil [52]	✓	✓	Otsu's threshold algorithm with ultrasonic sensor and RGB-D camera
Hsieh et al. [68]		✓	Using a CNN network

Huppert et al. [243]		✓	OptiTrack [244] environment with 12 cameras
Hussain et al. [177]		✓	Intel RealSense D435 depth camera
Islam et al. [245]		✓	Not mentioned in detail
Islam et al. [231]	✓		Ultrasonic sensor
Joshi et al. [130]	✓		Ultrasonic sensor
Kayukawa et al. [246]		✓	iPhone 11 pro camera
Kayukawa et al. [210]		✓	ZED RGB-D camera
Kayukawa et al. [247]	✓	✓	LiDAR sensor and RGB-D camera
Khairnar et al. [215]	✓		Ultrasonic sensor
Kim et al. [124]		✓	Distance estimation using a stereo camera
Kommey et al. [36]		✓	Based on the size of bounding boxes
Kuribayashi et al. [208]	✓	✓	LiDAR sensor and iPhone 11 pro camera
Li et al. [45]		✓	ZED RGB-D camera

Appendix 2 - Techniques used for obstacle detection

Requirements
1.Audio descriptions of high quality
2.Accuracy of directions and information
3.Alerts for unexpected events
4.Real-time performance for detection/recognition tasks
5.Ease of use, natural/intuitive user interface, acceptable by a broad user population, including senior citizens
6.A simple training procedure, potentially scalable to new objects and personalization
7.Tolerance to viewpoint variations
8.Tolerance to illumination variations

9.Tolerance to blur, motion blur, out of focus, and occlusions
10.Providing information on location, guidance and navigation
11.Providing information on the system operation
12.Selection of the device operation mode offline-online
13.Personalization by giving the ability to the users to define their own level of disability
14.Minimize the dangers and errors by preventing consequences of incidental or unintentional activity
15.Sharing information for accompanying contents of surroundings (coffee shops, hotels, hospitals, etc.)
16.A camera for detecting obstacles for also obstacle avoidance (moving and static objects/obstacles' shape, location, moving speed, etc.)
17.Recognizing the color of clothes
18.An affordable price
19.Early alert for obstacles, especially in a waist level
20.Position restore actions when the user gets lost
21.Notification of uneven floor surfaces such as loose street tiles, puddles or other small holes
22.Systems should reliably provide relevant information when needed, while also considering information accuracy
23.The types of obstacles that are communicated to the user should be restricted to those that are unexpected. This is, especially important to limit information overload and reduce system complexity
24.Different contexts may require different types of user interaction. Environments with many obstacles may require different types of notifications (i.e., more frequent, closer in range)
25.A better detection of horizontal objects, ground and small objects as well
26.Smaller and more efficient device
27.Obstacle recognition after detection (information in the output that would allow the user to distinguish between different types of obstacles)
28.A strong enough vibration signal to indicate an imminent collision
29.Detection of moving obstacles and small objects
30.Accurate voice and language recognition
31.Short sentences to be used as input configuration commands to the assistive mobile application
32.Combination of audio and touch model: Audio navigation commands and vibration alerts for early obstacle warning (2 m distance) with crescent frequency
33.Directions to avoid obstacles with vibration signals or audio guidance
34.Body wearable product (the preferred position is the waist)
35.System run locally on the device without the need for internet connection
36.Light product with small size
37.A clear description of the indoor place to create a mental map of it
38.Triggering/Sharpening the visually impaired user's environmental sensing

39.System should cooperate with human helper that guide the visual impaired

Appendix 3 - P-VI/blindness user requirements for scene understanding

Kind of assistance	**Paper**	**Year**
Obstacle detection	Martinez et al. [37], Caraiman et al. [118], Endo et al. Tapu et al. [240], Tapu et al. [248], Zhang and Ye [50], Srinivasan et al. [249], Malek et al. [129], van Erp et al. [250]	2017
	Wang et al. [38], Gandhi and Gandhi [114], Elmannai and Elleithy [74]	2018
	Hakim and Fadhil [52], Presti et al. [111], Rahman et al. [251], Canez et al. [51], Zhao et al. [200], Saha et al. [221], Guerreiro et al. [134], Li et al. [252], Megalingam et al. [253]	2019
	dos Santos et al. [239], Ntakolia et al. [108], Breve and Fischer [237], Hussain et al. [177], Khairnar et al. [215], Joshi et al. [130], Srinivasan et al. [249], Kandoth et al. [207], Calabrese et al. [131], Hsieh et al. [68], Gay et al. [241], Kayukawa et al. [246], Cheng et al. [234], Suny et al. [209], Islam et al. [231], Kayukawa et al. [210], Akula et al. [236], Kayukawa et al. [247]	2020
	Wang et al. [254], Chen et al. [238], Huppert et al. [243], Kuribayashi et al. [208]	2021
Object detection	Sosa-García and Odone [49], Lin et al. [72], Wang et al. [132], Silva and Wimalaratne [222], Tapu et al. [255], Dasila et al. [223], Rizzo et al. [256], Malek et al. [129], Kedia et al. [257], Jabnoun et al. [258]	2017
	Wang et al. [211], Jiang et al. [26], Gianani et al. [212], Dosi et al. [27], Eckert et al. [35], Liu et al. [112], Vyavahare and Habeeb [115], Hudec and Smutny [259], Mante and Weiland [133], Islam et al. [245], Arakeri et al. [116], Awad et al. [260]	2018
	Baskaran et al. [220], Rahman et al. [122], Duman et al. [34], Thomas et al. [120], Li et al. [45], Kim et al. [124], Kommey et al. [36], Suresh et al. [113], Tepelea et al. [226], Bharatia et al. [25], Guerreiro et al. [134],	2019

	Bazi et al. [261], Georgiadis et al. [235], Tahoun et al. [218], Saha et al. [221], Pehlivan et al. [125]	
	Afif et al. [227], Ghosh et al. [262], Yohannes et al. [202], Abraham et al. [205], Suresh et al. [230] , Vaidya et al. [213], Nguyen et al. [214], Chen et al. [206], Calabrese et al. [131], Shen et al. [151], Islam et al. [231], Shah et al. [216], Eskicioglu et al. [232], Fusco and Coughlan [263], Iwamura et al. [201], Ntakolia et al. [108], Alhichri et al. [126], Shelton and Ogunfunmi [228], Mandhala et al. [204], Hussain et al. [177], Aralikatti et al. [127], Bhole and Dhok [128], Khairnar et al. [215], Joshi et al. [130], Srinivasan et al. [249], Vaidya et al. [206], Imtiaz et al. [233], Cheng et al. [234], Suny et al. [209], Kayukawa et al. [210], Ahmetovic et al. [264], Boldu et al. [217]	2020
	Sarwar et al. [150], Kuribayashi et al. [111], Akkapusit and KO [139]	2021
Navigation	Endo et al. 2017 [240]	2017
	Li et al. 2018 [112], Elmannai and Elleithy [74]	2018
	Rahman et al. 2019 [251], Li et al. 2019 [252], Bazi et al. 2019 [261]	2019
	Khairnar et al. [215], Suresh et al. [230], Alhichri et al. [126], Suresh et al. [230], Gay et al. [241], Kayukawa et al. [210], Eskicioglu et al. [232], Fusco and Coughlan [263]	2020
	Chen et al. [238], Wang et al. [254]	2021
Face detection	Kedia et al. 2017 [257]	2017
	Hudec and Smutny 2018 [259]	2018

	Rahman et al. 2019 [122]	2019
	Aralikatti et al. 2020 [127], Alhichri et al. 2020 [126], Shah et al. 2020 [216], Ahmed et al. 2020 [265]	2020
	Sarwar et al. 2021 [266]	2021
Emergency calls	Vyavahare and Habeeb 2018 [66], Gandhi and Gandhi 2018 [114]	2018
	Suresh et al. 2019 [113], Bharatia et al. 2019 [25]	2019
Text recognition and reading	Caraiman et al. 2017 [118]	2017
	Arakeri et al. 2018 [116]	2018
	Thomas et al. 2019 [120]	2019
	Abraham et al. 2020 [205], Srinivasan et al. 2020 [249]	2020
Stop light detection and passing crossings	Li et al. 2019 [45]	2019
	Yohannes et al. 2020 [202], Chen et al. 2020 [214]	2020
Live location tracking	Suresh et al. 2019 [113], Bharatia et al. 2019 [25]	2019

Appendix 4 - Assistance services

Article	**Testing type**	**Subjects**	**Number of subjects**	**VI/Blind testing**
Ahmed et al. [265]	Controlled environment testing and Survey (NASA TLX form)	Blind and visually impaired	13	✓
Akkapusit and Ko [219]	Controlled environment testing, Survey (Likert-like scale questions), Interview	Blindfolded	20 (14 male and 6 female)	✓
Awad et al. [260]	Controlled environment testing, Survey (Likert-like scale questions)	Visually impaired	10	✓
Boldu et al. [217]	Controlled environment testing, Survey (Likert-like scale questions)	Blind	9 (8 male and 2 female)	✓
Canez et al. [51]	Controlled environment testing	N/A (Not available)	N/A	✓

Caraiman et al. [118]	Personal interview and survey	Visually impaired and blindfolded	24 (12 VI and 12 Sighted)	✓
dos Santos et al. [239]	Field experiments and Survey	Blindfolded, Blind	41 (10 blind and 31 blindfolded)	✓
Elmannai and Elleithy [74]	Field experiments (outdoor and indoor environments)	Blindfolded	1	✗
Eskicioglu et al. [232]	Controlled environment testing	Blind and Sighted	5 (3 sighted and 2 blind)	✓
Fusco and Coughlan [263]	Controlled environment testing	Blind	6	✓
Gandhi and Gandhi [114]	Controlled environment testing	N/A	N/A	N/A
Gay et al. [241]	Controlled environment testing	deaf-blindness	5	✓
Ghosh et al. [262]	Controlled environment testing, Survey (Likert-like scale questions)	Blindfolded	2	✗
Guerreiro et al. [134]	Field experiments and Survey	Blind	10 blind	✓
Huppert et al. [243]	Controlled environment testing	Blind and visually impaired	10 (4 female, 6 male)	✓
Hussain et al. [177]	Controlled environment testing	N/A	2	✗
Iwamura et al. [201]	Controlled environment testing, Interview and Survey (Likert-like scale questions)	Visually impaired	7	✓
Jiang et al. [26]	Survey (Likert-like scale questions)	Blind and Sighted	2 groups (number of participants is not mentioned)	✓
Joshi et al. [130]	Field experiments (outdoor and indoor environments)	N/A	36 (20 VI and 16 blindfolded)	✓
Kayukawa et al. [246]	Controlled environment testing, Survey (Likert-like scale questions)	Blind	6	✓
Kayukawa et al. [210]	Controlled environment testing, Survey (Likert-like scale questions)	Blind (4 male and 2 female)	6	✓
Kayukawa et al. [247]	Controlled environment testing, Field experiments and Survey (Likert-like scale questions)	Blind	14	✓
Kim et al. [124]	Controlled environment testing	Blindfolded	1	✗
Kuribayashi et al. [208]	Survey (Likert-like scale questions), Interview	Blindfolded, Blind	6	✓

Li et al. [252]	Controlled environment testing	Blindfolded, Blind	N/A	✓
Lin et al. [72]	Field experiments and Survey(Likert-like scale questions)	N/A	4	N/A
Liu et al. [112]	Remote usability testing(recording user's behavior and analyzing it later) and Think-aloud protocol	Blind	7	✓
Mante and Weiland [133]	Controlled environment testing (ANOVA analysis) and Interview	Early blind and late blind	12 (5 early blind, 7 late blind)	✓
Martinez et al. [37]	Surveys (NASA TLX form)	Blindfolded	6 (2 female and 4 male)	✗
Ntakolia et al. [108]	Controlled environment testing, Survey (Likert-like scale questions)	Blindfolded	10	✗
Presti et al. [111]	Think-aloud protocol and Survey (Likert-like scale questions including System Usability Scale (SUS))	Blind/visually impaired	13 (7 male and 6 female)	✓
Rahman et al. [251]	Field experiments	N/A (testers are called actors)	N/A	N/A
Saha et al. [221]	Controlled environment testing	Blind	13	✓
Shen et al. [151]	Controlled environment testing	Blind and Blindfolded	6 (3 blind and 3 blindfolded)	✓
Sosa-García and Odone [49]	Survey (Likert-like scale questions) and Field experiments	Blindfolded, blind and visually impaired	8 (4 blindfolded, 4 blind (1 congenitally blind and 3 low vision) 2 women and 6 men	✓
Stearns and Thieme 2018 [267]	Controlled environment testing	Blindfolded	5	✗
Suresh et al. [113]	Controlled environment testing	Blind	3	✓
Tapu et al. [248]	Field experiments	Visually impaired	N/A	✓
van Erp et al. [250]	Controlled environment testing, Survey (Likert-like scale questions	Visually impaired	5	✓
Wang et al. [132]	Survey (Likert-like scale questions)	Blind	15 (9 congenitally blind)	✓
Wang et al. [38]	Survey (Likert-like scale questions)	Blindfolded	6 (3 female, 3 male)	✗
Zhang and Ye [50]	Survey (Likert-like scale questions) and Controlled	Blindfolded	7	✗

	environment testing (indoor environment)			
Zhao et al. [200]	Controlled environment testing	Blindfolded	1	✗

Appendix 5 - Evaluation approaches

Questions about previous navigation systems experience	1. What tools are you currently using for navigation? What are the problems with those? 2. Have you ever used a digital assistance system before? Have you ever used an indoor version?
Questions about information modality and content	1. What unit of measurement do you usually use to measure or interpret distances (eg meters, steps,...)? 2. Present artificial scenario ([Known, unknown], [Bathroom, living room, kitchen, etc.]) How do you currently obtain information about the scenario? (Explore thought process) 3. Same scenario. What information would you like to receive to understand the environment better? And to locate an object in the environment? 4. How would you like to receive information about distances with objects? What method/format would be most useful for indicating distances (e.g. through sound, words, vibrations)? a. [Sound/Haptic]: Why? Ask him/her to provide some explanations. 5. Artificial scenario: moving user approaches an object. What information would you find helpful in locating the object? How and how often would you find it useful to receive the information?
Questions to explore user needs	1. What tasks are difficult for you because of your visual impairment? [Explore new use cases] 2. What tasks can a scene understanding system do to improve your life quality? (Focus on tasks that include object detection and depth estimation) 3. Describe an example. How would a recommender system help you? The system recommends actions based on your situation. a. What do you think of a functionality that notifies you about objects that may be interesting to you? For example, a record player with which you can listen to music. b. What do you think about a functionality that notifies you about objects/situations that may be dangerous for you? For example, a slippery floor or fragile objects.
Questions about scene description	1. Do you find the feature of scene understanding useful? 2. Entering a room for the first time. Would you like to know what items are present in your surroundings and/or know the type of room you are in? 3. In which situations is it useful to know the type of room you are in? 4. In which situations is it useful to know about the objects present in a room? a. Items: How would you want to receive information regarding the items?(the modality of information. for example, vibration/voice) b. Room: Any other things to say besides the type of room? 5. What kind of information do you need about the objects? (e.g. location, color or category of the object) How detailed?

	6. Would you find the item description feature for closer ranges useful? (ex. sitting at a desk) What kind of information would you like to get when you are sitting at a table with some objects on it?
Questions about object finding	1. Do you frequently need to look for an object during a day/week? [How often would the feature would be used] 2. What are the objects you look for the most? 3. How do you look for something? (a dynamic object/static object) How often do you encounter problems? Which ones? [Discover the thinking process for the use case and potential problems] 4. Would you use a digital system to look for a specific object? [Viability of use case]
Ending question	1. Is there anything else you would like to discuss?

Appendix 6 - User research questions

		Participant 1 (low vision, night blindness)	**Participant 2 (very low vision)**	**Participant 3 (blind - only light perception)**	**Participant 4 (congenitally blind)**
Existing experience	Navigation	white cane	white cane	white cane	white cane
	Scene understanding	Ask others, move around carefully	Ask others, move carefully with white cane	Ask others, move around carefully	Ask others, move around carefully
	Object finding	Ask others, move around carefully	Ask others, Aira	Ask others, move around carefully	Ask others, move around carefully
	Digital services	Google maps, screen reader	GPS apps, Braille display, voice over, be my eyes, Aira, Seeing AI	Seeing ai(barcode reader and OCR), BeMyEyes, Google maps walking mode, blind aquare	Nearby Explorer, Google maps walk mode
Modality	Beeps	✓	✓	X	✓
	Vibration	X (afraid of missing)	✓	X	✓
	Distance unit	Meter, Angle	Feet, Cardinal direction	feet, inch, clock method, using reference object	meters, clock face method, reference objects
	Augmented audio	unfamiliar	X (prefers verbal directions)	X	✓ (I have played augmented

					audio games and I think it can be useful.)
Scene understanding	Objects	✓moving objects and stuff on the floor	✓ Hierarchical information	✓ Stairs	✓ (Hierarchical information, windows, kitchen objects)
	Text reading	✓	✓	✓	✓
	Obstacle detection	✓(specially moving obstacles)	✓ (upper body obstacles)	✓ (obstacles like columns)	✓
	Scene type	N/A	✓	✓	✓
Object finding	High range	✓	✓	✓	✓
	Close range	✓	✓	✓ (close range is easier in general)	✓
Recommendation system	Hazards	✓	✓	✓	✓
	Interesting objects	X	✓	X	✓ (empty seats, vending machines, supermarket shelves)
Losing objects	Frequency	N/A	Rarely	Never (all household are blind and everything has its place)	Everyday (people change the objects places or I forget)
	Objects	Table 8	Table 8	X	Table 8
Most complex tasks		finding objects when others move them	dealing with images and audios with no description	finding stairs, finding registers in big stores	cooking, finding directions indoors
		Participant 5 (acquired blindness during high school)	**Participant 6 (congenital blindness)**	**Participant 7 (congenital blindness)**	**Participant 8 (lost vision at 14, right eye fully blind, left**

					eye 20 percent blind. Sensitive to light)
Existing experience	Navigation	Mostly guide dog	white cane	white cane	Wheelchair user
	Scene understanding	Ask others, move around carefully	Ask others, move around carefully	Ask others, move around carefully	Ask others for help
	Object finding	Using guide dog, move around carefully, ask others	Ask others, move around carefully	Ask others, move around carefully	Ask others for help
	Digital services	Seeing AI, Google maps, Compass app, Aira	dot walker, Google maps, Envision AI	Seeing AI, lookout, VoiceOver image description	iOS VoiceOver
Modality	Beeps	✓	✓	✓	✓
	Vibration	✓	✓ But preferred beep for more accuracy	✓	✓
	Distance unit	Steps	Meters	Meters, clock face method	preferred beeping over meters
	Augmented audio	X (never used)	X	X	X
Scene understanding	Objects	✓	✓	✓ Hierarchical information	✓
	Text reading	✓	✓	✓	✓
	Obstacle detection	✓	✓	NA	✓
	Scene type	✓	✓	✓	NA
Object finding	High range	✓	✓	✓	NA
	Close range	✓	✓	✓	NA
Recommendation system	Hazards	✓	X	NA	NA
	Interesting objects	X	✓	NA	NA
Losing objects	Frequency	Everyday	Everyday	Everyday	Sometimes

	Objects	Table 8	Table 8	Table 8	NA (due to limited movements)
Most complex tasks		Finding stuff in the new bathrooms/kitchens, cooking, using appliances, little steps in American houses	cooking, finding stuff in the new bathrooms/kitchen	moving in the streets	reading paper mails