



Analysis and design framework for the development of indoor scene understanding assistive solutions for the person with visual impairment/blindness

Moeen Valipoor¹ · Angélica de Antonio¹ · Julián Cabrera²

Received: 16 September 2023 / Accepted: 4 May 2024 / Published online: 18 May 2024
© The Author(s) 2024

Abstract

This paper discusses the challenges of the current state of computer vision-based indoor scene understanding assistive solutions for the person with visual impairment (P-VI)/blindness. It focuses on two main issues: the lack of user-centered approach in the development process and the lack of guidelines for the selection of appropriate technologies. First, it discusses the needs of users of an assistive solution through state-of-the-art analysis based on a previous systematic review of literature and commercial products and on semi-structured user interviews. Then it proposes an analysis and design framework to address these needs. Our paper presents a set of structured use cases that help to visualize and categorize the diverse real-world challenges faced by the P-VI/blindness in indoor settings, including scene description, object finding, color detection, obstacle avoidance and text reading across different contexts. Next, it details the functional and non-functional requirements to be fulfilled by indoor scene understanding assistive solutions and provides a reference architecture that helps to map the needs into solutions, identifying the components that are necessary to cover the different use cases and respond to the requirements. To further guide the development of the architecture components, the paper offers insights into various available technologies like depth cameras, object detection, segmentation algorithms and optical character recognition (OCR), to enable an informed selection of the most suitable technologies for the development of specific assistive solutions, based on aspects like effectiveness, price and computational cost. In conclusion, by systematically analyzing user needs and providing guidelines for technology selection, this research contributes to the development of more personalized and practical assistive solutions tailored to the unique challenges faced by the P-VI/blindness.

Keywords Assistive technology · Blindness · Visual impairment · Scene understanding

1 Introduction

There are around 285 [1] million people worldwide who have vision impairment. Unfortunately, they struggle a lot in doing daily tasks like finding their belongings, navigating, or avoiding obstacles.

During the past years, with the advances in deep learning and computer vision, more researchers have delved into the field of smart assistive solutions for the persons with visual impairment/blindness. These new solutions aim to overcome the limitations that traditional ones like white canes or guide dogs have. For example, white canes cannot be used to detect obstacles that are further away, and they are only useful for the detection of obstacles on the footpath that are not higher above the knees. Guide dogs, on the other hand, have more advantages in comparison with white canes, but they also have their limitations like the high costs for raising, training,

Communicated by Y. Zhang.

✉ Moeen Valipoor
moeen.valipoor@alumnos.upm.es

Angélica de Antonio
angelica.deantonio@upm.es

Julián Cabrera
julian.cabrera@upm.es

¹ Madrid HCI Lab, Universidad Politécnica de Madrid, ETS Ingenieros Informáticos, Madrid, Spain

² Grupo de Tratamiento de Imágenes Information Processing and Telecommunications Center, ETSI Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain

and taking care of the animal [2]. Both are useless for scene understanding, object location and other tasks.

In previous research [3], we performed a systematic mapping of the current computer vision-based assistive solutions for scene understanding. We discussed their main features and compared their advantages and disadvantages. Scene understanding for people who are blind is mainly focused on providing scene information that is not accessible to the user due to their disability. Getting to know the objects/obstacles/persons in their surroundings and locating their position can be a burdensome task for them, especially in an unfamiliar environment. The challenges of people who are blind extend to a wide range of tasks. Avoiding obstacles, detecting colors [4], reading notices, locating fallen or lost objects, and detecting faces are some of their main concerns [5]. Traditional assistive solutions like white canes and guide dogs can help to address these challenges to some extent, but their functionality remains very limited. For instance, as stated by Conradie et al. [6], people who are blind and used a cane and a dog, expressed a significant need to be notified of uneven floor surfaces like loose street tiles, puddles or other tiny holes, as well as the position of things that cannot be reliably located by a dog. In another study, Wang and Yu [7], demonstrated how daily activities of people who are blind are limited to a fairly small space due to lack of information about their surroundings because of their visual impairment.

Providing descriptions about a scene can improve the process of cognitive map formation in people with blindness which leads to a better spatial awareness [5]. Many researchers have tried to tackle this problem. Liu et al. [8] utilized augmented reality (AR) to develop a scene understanding feature in their solution that scans the environment and gives a quick overview of the scene. In their solution, each object in the environment has a voice and communicates with the user on command. Their system can support a wide range of visual cognition tasks, including navigation, obstacle avoidance, scene comprehension, and the creation and recall of spatial memories, with little training. Participants in the study showed a positive performance and feedback after testing the solution. Other researchers [9–11] developed similar scene understanding solutions and discussed the importance of such feature in assistive solutions. In [11] the developed solution uses a Raspberry Pi and a Night Vision Camera to classify scenes using image enhancement techniques which work in dark environments. Furthermore, in [9], researchers propose a solution which provides a list of objects present in the scene with a multiobject recognition scheme designed to recognize objects in large quantities in a given indoor scene. However, the focus of their study is on the presence or absence of objects rather than their position in the environment.

Regarding navigation assistance in [10] a solution is designed to help visually impaired individuals with four

modules: obstacle recognition, avoidance, indoor and outdoor navigation, and real-time location sharing. The proposed system combines a smart glove with a smartphone app that works well in low lighting conditions and detects obstacles in the users' surroundings. In a different study [12], researchers used planes and point clouds to determine the 3D location of obstacles using Apple's ARKit 2 [13] framework. Both the horizontal and vertical planes can be detected by ARKit 2. As such, it has the ability to distinguish between the ground and other planes that might pose a threat. In their approach, obstacles with more intricate shapes can be identified using point clouds, which are collections of points that depict the salient contours of objects in an image.

Additional solutions have been proposed for other challenges like obstacle avoidance, text recognition and face detection [14–16].

However, these solutions differ in several variables such as level of detail that the solution provides in the scene description, type of environment to which it can be applied (dynamics, room and objects size, obstacles, lighting conditions, etc.). These are criteria that need to be considered during the design process.

Moreover, there are some commercial solutions in the market for the person with visual impairment (P-VI)/blindness that try to address some of their challenges. For instance, mobile apps like Envision [17] 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 Seein-gAI [18] and Lookout [19] by Microsoft and Google respectively. [17]

Other apps, such as Aira [20] and Be My Eyes [21], 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 [22].

Be My Eyes has recently added a feature to the app called BeMyAI [21] that uses the power of multimodal large language models (MLLM) which supplement text-based knowledge with additional data modalities such as visual data [23]. 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 [23].

Modern assistive solutions have the potential of a significant improvement in the lives of persons with visual impairment/blindness, but for this potential to be realized there are still important issues that need to be addressed. This paper concentrates on two of these issues: lack of user-centered approach in solution development, and lack of guidelines for the selection of appropriate technologies. We focus on scene understanding assistive technologies, however, similar gaps can be found in other assistive technologies and the proposed approach could be extended to them in the future.

2 User-centered design and development

It is critical to learn about the needs of users with a VI/blindness in order to design a usable assistive solution. In recent years, several assistive solutions have been designed without careful evaluation of user needs and priorities.

In our previous research [3], we gathered and analyzed the solutions that had been developed in this field, and among the gaps that we found were the lack of a deep understanding and consideration of the needs of P-VI/blindness and the absence of sufficient user testing with the target population. In our 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.

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.

There is not much assistive technologies research that focuses especially on user understanding and requirements elicitation with users with VI/blindness. Wise et al. [24] evaluate the effectiveness of various indoor navigation solutions for the people who are blind and discuss 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. 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 [25], 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. Researchers in [26] conducted a survey 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 less interest, while the ones with retinitis pigmentosa (RP) and glaucoma (GI) only required it for peripheral vision.

Their work demonstrates why such solutions must be customizable and why 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 a study by Conradie et al. [6] 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 [27] 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, some of the proposed solutions may not be very useful for the end user as a result of neglecting their needs, concerns, and the range of possible situations in the intended context of use.

In the following section 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.

Our 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 [28]. 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 [29]. 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 [30]. This implies that a new set of solutions and technologies for understanding indoor scenes must be developed.

Our work is 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. Given these considerations, the use cases we define in Sect. 2.2 are tailored specifically for indoor environments. These scenarios are intended to address the unique challenges associated with understanding indoor scenes and to discuss practical, usable solutions for people with a visual impairment. They are defined based on our previous systematic analysis of literature [3] as well as our primary user research which focuses on several challenges that people with a VI/blindness face in their daily lives regarding indoor spatial awareness.

2.1 User research

Our user research began with an examination of previous initiatives of other researchers aimed at understanding the needs of the community [3]. The main themes found for scene understanding in assistive solutions were related to object recognition, obstacle detection, object localization, scene type recognition and text reading.

However, to verify these findings, and figure out unexplored user needs for developing a solution specifically for indoor scene understanding, we carried out a series of semi-structured interviews with people that had different levels of visual impairment and blindness. Eight participants were selected who live in different parts of the world. Interviews were aimed to find out the ways that people with a VI/blindness use to undertake their daily tasks and overcome their challenges, and to discover the needs of the people with blindness in indoor environments regarding scene understanding.

Later in the use cases (Sect. 2.2), it will be explained how user needs would be addressed. After that, in Sect. 3, it will be discussed how these requirements were addressed in previous studies and the associated challenges.

2.1.1 Methods

According to our previous study, a list of questions (Table 1—Semi-structured Interview questions) was prepared for the semi-structured interviews to explore more in depth the area of scene understanding. The participants were found through social media (mainly Twitter and Instagram). The participants had a diverse range of visual impairments, including congenital blindness, light perception, night blindness or acquired blindness at different stages of their life. All of the interviews were held online through Zoom or Google Meet. The sessions were recorded with the consent of the user and later analyzed to discover all the important details. The language of the interviews were Farsi and English since interviewees were from English and Farsi speaking countries such as the UK, USA and Iran. Once the information was obtained from the interviews, a thematic analysis was applied to extract the common points that were mentioned by various users. The coded information obtained from the interviews is in Table 2—Interview results.

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 and the prepared questions were asked. Additionally, during the interview, some hypothetical situations (e.g. entering a new unfamiliar hotel room or sitting at a table in a restaurant) were presented and the user responses were analysed considering their level of disability to explore the various user needs and figure out how assistive solutions can assist them in various situations.

The main topics covered in the interviews, though not all of them were relevant to all participants, were as follows: 1. The actual use of current assistive solutions, 2. Scene understanding, orientation and mobility in different environments and the formation of mental maps, 3. Different output modalities of the assistive solutions, 4. Most difficult tasks they undertake in their daily lives, 5. Frequency and the kind of objects they lose, 6. Recommendation systems that could inform them about hazards or interesting objects in the environment.

2.1.2 Results

The findings are organized into sections based on key themes identified during the initial thematic coding of the results.

2.1.2.1 Existing assistive solutions In the case of questions regarding the 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, some users employed digital assistive

Table 1 Semi-structured interview questions

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?

solutions such as AIRA [20], Google Maps, Seeing AI [18], Be My Eyes [21], 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. 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!" Another participant expressed discomfort in using solutions such as AIRA or Be my eyes in which an actual person assists your through 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

Table 2 Interview results

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	White cane	White cane
Modality	Scene understanding	Ask others, move around carefully	Ask others, move around carefully with white cane	Ask others, move around carefully	Ask others, move around carefully	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	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	Seeing ai(barcode reader and OCR), BeMyEyes, Google maps walking mode, blind aquare	Nearby explorer, google maps walk mode	Nearby explorer, google maps walk mode
	Beeps	✓	✓	X	X	✓	✓
Scene understanding	Vibration	X (afraid of missing)	✓	X	X	✓	✓
	Distance unit	Meter, angle	Feet, Cardinal direction	Feet, inch, clock method, using reference object	Feet, inch, clock method, using reference object	Meters, clock face method, reference objects	Meters, clock face method, reference objects
	Augmented audio	unfamiliar	X (prefers verbal directions)	X	X	✓ (I have played augmented audio games and I think it can be useful.)	✓ (I have played augmented audio games and I think it can be useful.)
	Objects	✓ Moving objects and stuff on the floor	✓ Hierarchical information	✓ Stairs	✓ Stairs	✓ (Hierarchical information, windows, kitchen objects)	✓ (Hierarchical information, windows, kitchen objects)
Object finding	Text reading	✓	✓	✓	✓	✓	✓
	Obstacle detection	✓ (Specially moving obstacles)	✓ (Upper body obstacles)	✓ (Obstacles like columns)	✓ (Obstacles like columns)	✓	✓
	Scene type	N/A	✓	✓	✓	✓	✓
	High range	✓	✓	✓	✓	✓	✓
Recommendation system	Close range	✓	✓	✓ (Close range is easier in general)	✓ (Close range is easier in general)	✓	✓
	Hazards	✓	✓	✓	✓	✓	✓
	Interesting objects	X	✓	X	X	✓ (Empty seats, vending machines, supermarket shelves)	✓ (Empty seats, vending machines, supermarket shelves)
	Frequency	N/A	Rarely	Never (all household are blind and everything has its place)	Never (all household are blind and everything has its place)	Everyday (people change the objects places or I forget)	Everyday (people change the objects places or I forget)
Losing objects	Objects	Lost objects table	Lost objects table	X	X	Lost objects table	Lost objects table
	Most complex tasks	Finding objects when others move them	Dealing with images and audios with no description	Finding stairs, finding registers in big stores	Finding stairs, finding registers in big stores	Cooking, finding directions indoors	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	White cane	Wheelchair user	Wheelchair user
	Scene understanding	Ask others, move around carefully	Ask others, move around carefully	Ask others, move around carefully	Ask others, move around carefully	Ask others for help	Ask others for help

Table 2 (continued)

	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)
Modality	Object finding	Using guide dog, move around carefully, ask others	Ask others, move around carefully	Ask others for help
	Digital services	Seeing AI, Google maps, Compass app, Aira	Dot walker, Google maps, Envision AI	IOS VoiceOver
	Beeps	✓	✓	✓
Scene understanding	Vibration	✓	✓ But preferred beep for more accuracy	✓
	Distance unit	Steps	Meters	preferred beeping over meters
	Augmented audio	X (never used)	X	X
Object finding	Objects	✓	✓	✓
	Text reading	✓	✓ Hierarchical information	✓
	Obstacle detection	✓	✓	✓
Recommendation system	Scene type	✓	NA	NA
	High range	✓	✓	NA
	Close range	✓	✓	NA
Losing objects	Hazards	✓	X	NA
	Interesting objects	X	✓	NA
	Frequency	Everyday	Everyday	Sometimes
Most complex tasks	Objects	Table 5—Most frequent lost objects	Table 5—Most frequent lost objects	NA (Due to limited movements)
	Finding stuff in the new bathrooms/kitchens, cooking, using appliances, little steps in American houses	Finding stuff in the new bathrooms/kitchens, cooking, using appliances, little steps in American houses	Cooking, finding stuff in the new bathrooms/kitchen	Reading paper mails

actual person services it's just my anxiety over that. I prefer to ask someone around or just use internet."

2.1.2.2 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 units of measurement 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.

2.1.2.3 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."*

2.1.2.4 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. Two of the interviewees, P#5 and P#3, did not support the idea of having a recommendation system for the interesting objects.

2.1.2.5 Losing objects Participants were asked how frequently they missed items in their daily lives. Those who lived with sighted households 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 3—Most frequent lost objects.

2.1.2.6 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

Table 3 Most frequent lost objects

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

kitchens (P#5), reading mails (p#8) etc., which is mentioned in Table 2—Interview results.

2.2 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 [3]. 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 which the user with 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.

2.2.1 Scene understanding use cases

2.2.1.1 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 m 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 types 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).

2.2.1.2 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 m 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.

2.2.2 Object location use cases

2.2.2.1 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.

2.2.2.2 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 objects 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.

2.2.3 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.

2.2.4 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 [31]. 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 [32, 33] 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 4—Use cases.

2.3 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. The requirements are divided into functional and non-functional. Functional requirements are the capabilities, behavior, and the information that the system requires [34]. On the other hand, non-functional requirements focus on quality constraints like usability, reliability, availability, performance etc. [35].

2.3.1 Functional requirements

The functional requirements are 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. (Situation 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).

2.3.2 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.

Table 4 Use cases

	User need	Location	User position	Object/s	Object size	Distance
Use case 1	Scene recognition	Room	Standing	Static	Small & Big	> 1 m
Use case 2	Scene description	Room	Standing	Static	Small & Big	> 1 m
Use case 3	Scene description	Room	Standing	Dynamic	Small & Big	> 1 m
Use case 4	Scene description	Room	Moving	Static	Small & Big	> 1 m
Use case 5	Scene description	Room	Moving	Dynamic	Small & Big	> 1 m
Use case 6	Scene description	Room	Moving/Standing	Static/Dynamic duplicated objects	Small & Big	> 1 m
Use case 7	Scene description	Table	Sitting	Static	Small	< 1 m
Use case 8	Scene description	Table	Sitting	Static specific object	Small	< 1 m
Use case 9	Object location	Room	Standing	Static	Big	> 1 m
Use case 10	Object location	Room	Moving	Static	Big	> 1 m
Use case 11	Object location	Room	Moving	Static	Small	> 1 m
Use case 12	Object location	Room	Moving	Static	Big	> 1 m
Use case 13	Object location	Room	Standing	Static	Small	> 1 m
Use case 14	Object location	Room	Standing	Moving	Small	> 1 m
Use case 15	Object location	Table	Sitting	Static	Small	< 1 m
Use case 16	Color detection	Table	Sitting	Static	Small	< 1 m
Use case 17	Color detection	Table	Sitting	Static (any object)	Small	< 1 m
Use case 18	Obstacle avoidance	Room	Moving	Static	Small	< 1 m
Use case 19	Obstacle avoidance	Room	Moving	Dynamic	Small	< 1 m
Use case 20	Obstacle avoidance	Room	Moving	Static	Big	< 1 m
Use case 21	Obstacle avoidance	Room	Moving	Dynamic	Big	< 1 m
Use case 22	Hazard warning	Room	Moving	Static	Small/big	< or > 1 m
Use case 23	Text detection	Room	Standing	Static	NA	< 1 m
Use case 24	Text reading	Room	Standing	Static	NA	< 1 m

According to the discussion of researchers in [24, 34, 36] with several associations of persons with a 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. Reasons for considering the assistive solution “portable” are discussed later in the “User experience” section.

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 Sect. 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. Researchers in [27] 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 [37].

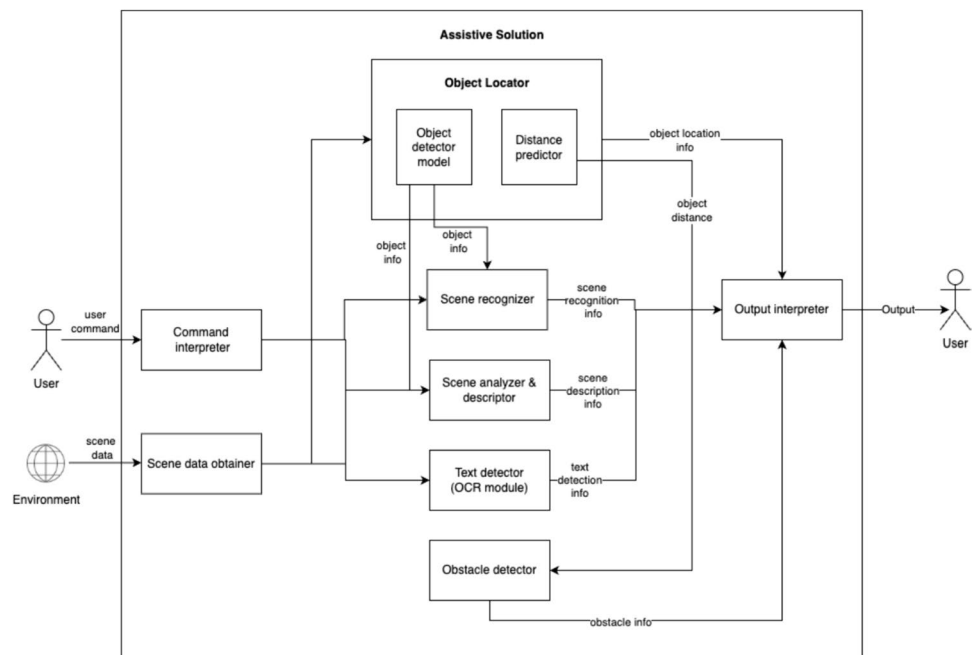
2.3.3 Reference system architecture

According to the functional and non-functional requirements, a high-level reference architecture for scene understanding assistive solutions is proposed which is presented in Fig. 1 – Solution Architecture: The figure shows different modules of the solutions and how they interact with each other. 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

Fig. 1 Solution Architecture: The figure shows different modules of the solutions and how they interact with each other



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 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.

2.4 User experience

In addition to functional and non-functional requirements for the implementation of the solution, the quality of the user's experience with such a system is also crucial. Designing a good user interface for users who are blind is challenging due to their sensorial limitations, different levels of disability and their different form of interaction with portable smart devices in comparison with sighted users.

Various studies showed that people with blindness would like to be able to interact with the assistive solutions through voice commands [34] because touch screens are not very efficient for them, and they cannot interact with traditional touch user interfaces as easily as a sighted person [38]. 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.

Example, a VUI for the user who is blind could be complemented with haptic feedback to communicate with the user [24]. Moreover, due to the aforementioned challenges, it's beneficial to provide a touch-based interface despite all the challenges that people with blindness have with the

dynamics and orientations of such interfaces [39]. This is because touch user interfaces provide more privacy to the user while retaining less of the user's hearing ability.

Personalization is also one of the aspects of assistive solutions that were discussed in [34]. 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.

This is because, as we have mentioned earlier, users with 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.

Besides the user experience aspects of the software solution, the hardware is also important. 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 [6]. 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 [6].

Furthermore, a significant advantage of 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 [40] suggest that the aesthetics of assistive technology and social acceptance may contribute to stigmatization of users. In a study by dos Santos et al. [41] researchers found that aesthetics seems to have a great influence on how people see and judge users. After interviewing eight participants with VI/blindness they found out that wearing modern devices that are subtle and have modern aesthetics (e.g. smart glasses) was more accepted by the participants. Therefore, it is an important factor that influences device adoption or abandonment.

3 Guidelines for the selection of appropriate technologies

After careful consideration of earlier work and the conclusions of our own prior study [3], it became evident that there is a variety of different approaches and technologies for the design of an assistive solution, with different features, cost, and performance.

When it comes to the development of a specific assistive solution, the selection of the most appropriate technologies and the adaptation of the system to the user needs imply non-trivial design decisions. The variety of technologies available for scene understanding, with many aspects to be considered like the algorithms used, the required hardware (e.g. camera and sensors), the precision, reliability or performance of the solution, cost, ergonomics, and others, make the selection of the best option a challenging task.

In this section, different technologies that can be used for different features of assistive solutions (scene recognition and description, object location, obstacle avoidance, and text detection) and modules of the high-level architecture are discussed.

3.1 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 [42]

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 [43]. In the use cases defined previously, use case 1 can be implemented using a scene recognition technique.

Early scene recognition methods were mainly based on global attribute descriptors which are composed of a few basic visual characteristics to simulate human perception [42]. 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 [44]. 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, the inference time for deep neural networks is significantly increased by a high number of sampled image patches. 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 [44].

Recently, with the advancements in deep learning and convolutional neural networks (CNN), more developed methods such as the ones in [45–47] were presented with a high classification performance. For scene recognition using CNNs, features from intermediate and high layers (e.g., parts and objects) 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 [44].

Recently, 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 [47] 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, UIUC Sports-8, MIT Indoor-67, SUN-397, Places, that can be used for the training of the scene recognition algorithms. Among the mentioned datasets, MIT Indoor-67 is focused on indoor environments making it a more suitable dataset for our use cases.

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. Researchers in [42] a scene recognition solution for the people with VI/blindness that is based on a EfficientNet [48] 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 [49], and Inception [30], can be trained for scene recognition on mobile devices.

3.2 Scene description

Individuals with visual impairments have a difficult time orienting themselves in various indoor/outdoor environments. Due to their visual limitations, forming a mental map of the space is a challenging task for them. Cognitive or mental maps are mental models for learning,

understanding, simplifying, and explaining human interactions with the environment, which include object locations, observations, routes, spatial relations, etc. [50]. Different solutions have been developed in the past years that tackle this challenge [9, 51, 52]. For instance, providing information about their surroundings in the form of audio or tactile feedback can be helpful for them in forming mental maps [53].

Moreover, by analyzing and describing image content from the users' surroundings, scene description 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. The process becomes even more challenging when describing a scene or an image to a person who is blind. It is 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.

For example, in Use case 2, the user is static and is interested in static objects that are more than 1 m away. In such conditions, it is probably more important to inform the user first about the existence and location of the largest objects that are closest to the user. For example, it is more important to know about the existence of a table first and then about the smaller objects placed on it. However, in Use cases 3, 4, and 5, where either the user or the objects are dynamic, it is more crucial to know about moving objects first to avoid collisions. Therefore, it is recommended to provide information in a hierarchical format to simplify the scene description process (this was also mentioned by participants in our interviews). For instance, naming the objects based on their status (static/dynamic), color, their relationships (e.g. "a red vase is on the table") and distance would be more informative for the user than generating a general phrase (e.g., "A vase inside a living room.") for scene description, which may not provide the information that the user with blindness needs. In the next sections the main components (object detection and object location) that are needed by the scene description module are discussed.

3.3 Object detection

Deep neural networks (DNN) for object detection, particularly the CNN models like ResNet [54] have significantly increased the potential of computer vision for creating assistive technologies for the persons with VI/blindness over the

past few years. Compared to shallow networks like AlexNet, they perform noticeably better, making real-time object recognition more feasible [55]. Additionally, CNNs are capable of automatically learning high-level semantic features from the input data, optimizing multiple tasks at once, including bounding box regression and image classification [56].

Various object detection models and algorithms such as YOLO, Single Shot Detector (SSD) and Inception-v3 [57] have been used in assistive solutions during the past years and were analyzed in our previous work [3].

Bhumbla et al. [58] compared object detection algorithms such as various versions of YOLO, SSD and R-CNN [59]. 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 [60, 61]. Therefore, it is important to choose a model that matches the use case of the assistive solution.

In Use cases 2, 3, 7, and 8, where the user is not moving, various object detection methods such as YOLO [49] could be used for detecting objects. YOLO is a CNN-based object detection method that makes use of Darknet, an open-source network framework written in C and CUDA. YOLO creates bounding boxes for each SxS grid it creates from an image. The probability of classes for objects and their corresponding bounding boxes is then predicted. In the case of Use cases 4, 5, the process is more challenging since the user is moving and camera movement can decrease the accuracy of the object detection algorithms. To tackle this problem, it is necessary for the system to have a very low response time, ensuring immediate feedback. This involves processing visual data, recognizing and tracking objects, interpreting changes in their positions and states, and then promptly generating up-to-date descriptions of the scene. To meet these demands, both the object detection, depth detection and natural language generation components of the system need to be optimized for speed, without compromising accuracy. For instance, some variations of YOLO could be used, like Tiny-YOLO which is made for mobile devices and recognizes a lower number of object classes, but it operates faster.

In Use cases 7, 8 and 14 in which small objects (such as a fork, spoon, or knife) are close to the user, object detection algorithms may not function effectively. This is because they only draw a rectangle around the objects and do not indicate the exact pixels where the object is located. This lack of information can lead to difficulties in distinguishing and describing small, closely placed objects like utensils, which could potentially create confusion for the user. For larger objects, the shortcomings of traditional object detection algorithms are less impactful

because these objects typically cover a large enough area in the image, making their presence and general location easy to identify with the bounding boxes. In the case of small objects, semantic and instance segmentation techniques are better to be used for locating the exact position of the object with respect to the user, because semantic segmentation provides more precise details regarding the existing objects in the scene compared with object detection bounding boxes. Each pixel in the scene is assigned a category to identify the object type it belongs to. Each pixel of the image is therefore given a semantic marking. Semantic segmentation identifies both the type of object and its location in the scene. Moreover, instance segmentation provides even more details. It is distinct from semantic segmentation in that each segmented object in an image is given a unique identifier. Segmentation of an instance will result from semantically segmenting an image with observed boundaries of each object [62]. This approach will also be useful for color detection in Use case 16 and 17, because the color of objects can be obtained by getting the HSV (hue, saturation and value) and RGB (red, green and blue) values of a pixel and turning it into natural language [63, 64]. 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.

Furthermore, object detection models can also be utilized as the basis for detecting potential hazards and obstacles around the user, such as in Use cases 20, 21 and 22. For instance, in [65] a hazard detection custom model of YOLOv4 is proposed for detecting high-voltage rooms, poison gas rooms, engine rooms, high terraces, doors leading to outdoors, and highways. In another study [66] researchers developed a guide belt for VI that uses YOLOv7 for pothole detection.

It is also possible to run and train all the mentioned models on a remote server in the cloud or use services such as Google Cloud Vision [67], Microsoft Azure Computer Vision [68], Amazon Rekognition [69], and so on to undertake the process remotely. In the case of the cloud services mentioned, they have been trained on massive datasets, allowing for improved performance. However, these solutions need constant connection to the remote server and lead to more power usage. Additionally, most of these service providers are not free and include costs per request process.

3.4 Object location–distance prediction

The distance predictor module of our reference architecture is a critical component for assisting people with a visual impairment in navigating the real world. This module should

determine the spatial location of objects in the physical environment. It works by leveraging advanced depth estimation technologies to determine the distance between the user and various objects around them. The process of detecting the location of objects in a real 3D environment is divided into two tasks: detecting the object in the scene and finding its location. For the first task, the object detection algorithms that have already been brought up for discussion in the previous section can be used. Nevertheless, determining the three-dimensional position of the object in the surrounding environment is a challenging task. There are several ways to complete this task, which will be discussed in the following sections.

3.4.1 Depth estimation based on depth cameras

Depth cameras are used to provide information about the distance between objects in a scene and the camera. Different camera technologies have been developed for this task, each type with its own benefits and drawbacks, which depend on the processing power, computational time, range limitation, cost, and the environment in which they work properly (e.g. lighting conditions, environment size, colors etc.).

3.4.1.1 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 [70]. 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 [71]. Nevertheless, 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 inaccuracies in distance 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.

In addition, some stereo cameras (such as the Intel RealSense D455) include a projector (an infrared (IR) projector that projects a dot pattern). The projection is used in this method to simplify the matching of camera images and has a significant advantage in low-light environments or when the texture is not very distinct.

The detection range of stereo cameras is determined by several factors, including the cameras' resolution, the baseline distance between them, the focal length of the lenses, and the accuracy of the disparity estimation algorithm [72]. In the case of Intel RealSense cameras, the range can vary from 0.1 to 10 m depending on the model and the settings.

3.4.1.2 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 Camara Intel RealSense SR305 [73]).

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 a study by Maculotti et al. [74] their findings demonstrate 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. Such cameras do not work well on highly reflective, transparent, very dark surfaces or in outdoor environments where the lighting is intense [75].

3.4.1.3 Time-of-flight (ToF) cameras Time-of-Flight (ToF) cameras (e.g. Kinect Azure [76]) 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 5.46 m [77]. 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 m [78]. 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 [79].

3.4.2 Monocular depth estimation methods

Besides the discussed methods which rely on depth cameras, there are some other approaches that are able to estimate depth distance from a single RGB image.

3.4.2.1 Deep learning-based monocular depth estimation Deep learning-based methods became popular during recent years. It has been demonstrated by the current state of the art that monocular depth estimation methods are a viable solution to various depth-related challenges. 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 [80]. 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 nighttime [81]. To enhance the accuracy of depth estimation, it is necessary to collect more data of higher quality and from a wider variety of scenes. The existing datasets used for depth estimation are limited, and creating a new dataset is both time-consuming and costly [82].

3.4.2.2 Mathematical-based monocular depth estimation methods There are also some mathematical methods that estimate the distance of objects in front of a monocular camera. Researchers in [83] used a method proposed by Davison et al. [84] for object distance estimation. They use camera height, center coordinate of the formed image, and distance from the camera to the target object's coordinates against the ground to calculate the distance of the object from the camera. However, this method has some limitations. For instance, it does not consider the actual size of the object. This means that the system cannot differentiate between a large object far away and a small object close by, which could both appear to be the same size in the image. Additionally, this method only provides the distance to an object's center which might not be accurate enough for some objects that have a high range of depth (e.g. a large sofa that is placed alongside the user).

Overall, stereo vision, structured light and time-of-flight cameras are the most promising options for object location use cases due to their accuracy. However, monocular depth estimation approaches are cheaper and more energy efficient. They are also a better choice for deployment on smaller assistive devices such as a smartphone since they only need a single camera.

3.5 Obstacle avoidance

One of the biggest daily challenges of people with blindness is obstacle avoidance. Especially the obstacles that are not at knee level (e.g. low hanging branches, overhangs, or other objects that are positioned above head level) cannot be detected using white canes.

There are different ways to detect obstacles in the environment. For instance, LiDAR sensors already discussed can be used to generate a 3D map of the surrounding environment. This information can then be used to identify obstacles (3D points at a distance lower than a threshold) and provide audio or tactile feedback to the user. For instance, researchers in [85], developed a LiDAR-based obstacle detection solution that works with a LiDAR camera connected to a Raspberry Pi. Their method shows promising results (average accuracy of 88%).

Furthermore, ultrasonic sensors use sound waves to measure the distance to objects by sending these waves into air and then using the reflected waves to measure distance. These sensors 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 [3], working in dim lighting, detecting transparent objects and low power consumption. Many solutions such as [86–88] were developed utilizing these sensors. However, there are a few drawbacks, including a short detection range (about 2 m) [89] and poor accuracy when detecting soft or curved objects. Moreover, ultrasonic sensors are only useful for obstacle avoidance but cannot be used for more complex tasks such as scene recognition or description, given that they are not able to associate distance information to specific objects.

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 to 150 cm [88]) is quite limited.

Camera-based techniques, in addition to sensors, can be employed for obstacle detection. Some researchers [90] have employed camera-based techniques to identify obstacles, either alone or in tandem with sensors. The range these cameras cover for detecting obstacles is usually between 5–7 m.

Additionally, in a couple of studies [15, 91] the Simultaneous localization and mapping (SLAM) method is used for navigation and obstacle avoidance. SLAM [92] can determine the position and orientation of the sensor relative to the surrounding environment while also mapping the environment. Visual SLAM [93] is a type of SLAM that uses camera image input to perform SLAM positioning and mapping in unknown environments. Researchers in [94] proposed a blind guidance system using both tactile and auditory feedback by combining ORB-SLAM [95] 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.

A more thorough discussion of various techniques and the methods used in existing assistive solutions can be found in our previous work [3]. The choice of the most suitable obstacle detection technique depends on the use case. For instance, ToF sensors such as LiDAR sensors would be a suitable option for moving objects (Use cases 16 and 18) since they have a fast response time [96], allowing for quick and accurate detection and tracking of dynamic obstacles. However, for more energy and price efficient approaches ultrasonic/infrared sensors could be used.

3.6 Text detection

Scene understanding involves being able to read the texts that are present around us. 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 be then read aloud by a text-to-speech software. There are some current solutions such as SeeingAI [18] 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 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].

3.7 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, including the computational resources required, cost, portability, simplicity, price, speed, battery life, aesthetics, and feedback type [33, 100, 101]

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.

To help designers in the selection process, we compared various strategies for implementing assistive solutions considering various aspects such as effectiveness, computational resources, limitations, and acquisition cost of each approach.

Complete information for the comparison of different implementation strategies is in Table 5—Solutions comparison. For each of these approaches we found solutions that were implemented using the algorithms and devices mentioned, to make the comparison more objective. However, for a few of the approaches we could not find information such as the computational resources, or effectiveness of the method in the corresponding existing solution or study.

3.7.1 Effectiveness

In the case of effectiveness, we compared the accuracy/average precision (AP) of different algorithms in combination with the detection range 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. However, instance segmentation is a better approach for finding the exact position of an object in an image.

For distance detection, 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 m 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 performance in terms of range in comparison with other sensors. The Realsense LiDAR L515 can detect obstacles up to 6 m away from the user. The ultrasonic and infrared sensors are also acceptable options for detecting close-range obstacles in cost-efficient solutions.

3.7.2 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 precision and performance will result in a higher price. For instance, for 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. For the implementation, we considered embedded devices. 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 performance of a tiny model is less 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).

3.7.3 Computational resources

High accuracy 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 research includes technological selection 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 quality 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.

3.7.4 Limitations

As we discussed in the previous sections, each of the mentioned approaches have their limitations, such as the precision and the computational resources they need for running. It is also important to note that the generalizability of proposed solutions might vary in different situations.

Table 5 Solutions comparison

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)) [104]	85.40% accuracy on indoor-67 dataset [44]	Not available	Only limited scenes	Not available
Scene recognition	Use case 1	CNN/Hybrid models (e.g. FOSNet [47])	90.37% accuracy on indoor-67 dataset [44]	Not available	Need more computational resources	Not available
Scene recognition	Use case 1	Lightmodels (e.g. SceneRecog [42])	83.33% [42] 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) [105]	Jetson Nano/ Raspberry Pi 4 according to [106]	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 [105]	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 [107])	43.7% AP on COCO dataset [108]	Jetson Nano according to [109]	Computationally heavy	Jetson Nano From 150\$ + Pi Camera 30\$
Scene description	Use case 6,7	Object detection (e.g. YOLOv7)	56.8% AP [105]	Jetson Nano/ Raspberry Pi 4 [106]	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 [21]	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 [110]	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 [105] for object detection 0.1 m to 4 m camera range [111]	Raspberry Pi 4 [106]	Not detecting textureless surfaces	Raspberry Pi 4 from 35\$ Realsense D455 is 419\$
Object location	Use case 8 to 17 and 22	Object detection (YOLOv7-tiny) + ToF camera (e.g. Intel RealSense LiDAR L515)	35.2% AP [105] for object detection 0.35 m to 6 m camera range [78]	Raspberry Pi 4 [106]	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 [112] v2.1 small)	35.2% AP [105] for object detection 13.43% mean absolute relative error (MARE) for MiDAS v2.1 [113]	Android Smartphone according to [114]	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 [105] for object detection 0.5 m camera range	Raspberry Pi 4 [106]	Requires more computational power than stereo camera	Intel RealSense SR305 410\$

Table 5 (continued)

User need	Use case	Approach	Effectiveness (Accuracy and camera range)	Computational resources	Limitations	Acquisition cost
Object location	Use case 8 to 17 and 22	Instance segmentation (e.g. Detectron2 R101-FPN model [107]) + stereo camera (e.g. Realsense D455)	43.7% AP on COCO dataset 0.1 m to 4 m camera range	Jetson Nano according to [109]	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 [107]) + ToF camera (e.g. Intel RealSense LiDAR L515)	43.7% AP on COCO dataset [108] for instance segmentation 0.35 m to 6 m camera range [78]	Jetson Nano according to [109]	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 [107]) + 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 to 6 m camera range [78]	Jetson Nano according to [109] or Rasp berry Pi according to [85]	Computationally heavy	Raspberry Pi 4 from 35\$ Realsense LiDAR L515 is 880\$
Obstacle avoidance	Use case 18 to 22	Ultrasonic sensor (e.g. HC-SR04)	Up to 0.4 m [115]	Rasp berry Pi according to [115]	Poor accuracy detecting soft or curved objects	Raspberry Pi 4 from 35\$ Ultrasonic sensor 8\$
Obstacle avoidance	Use case 18 to 22	Infrared sensor [116]	Around 20 to 150 cm [88]	Rasp berry Pi according to [88]	Short range (up to 150 cm)	Raspberry Pi 4 from 35\$ Infrared sensor 10\$
Obstacle avoidance	Use case 18 to 22	RGB-D Camera (Intel Realsense D415 [117])	Up to 2 m [117]	Jetson Nano according to [118]	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 [119]	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\$

The methods discussed in Table 2—Solutions comparison 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. Finally, user experience is a crucial part of assistive technologies that has received insufficient attention in many studies. The experience of the people with blindness varies greatly depending on their level of disability. For example, there is a substantial difference between the experiences of users who are totally blind, deaf blind, or with a visual impairment. In previous sections, we discussed the need for personalization of the solution and how it might be adapted to the user's requirements.

4 Discussion

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 developed/evaluated according to the mentioned framework, an existing solution (Seeing AI) has been assessed according to the proposed elements. We chose Seeing AI 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. 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. It does not include detecting specific objects (FR 4) although 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 is one of the power points of this app. There are options for "short text", "document" and "handwriting" that are included in the app. 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. Additionally, 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. 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.

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, researchers in [102] evaluated the performance of the text detection (OCR) functionality of this app 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 [103] Seeing AI was used for image description, and it was compared with two other solutions (Image Explore 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.

5 Conclusion

In this paper the main needs and challenges for developing scene understanding assistive solutions for the P-VI/blindness were identified through semi-structured interviews and analysis of the state of the art. According to the information obtained, we proposed a set of structured use cases that help to visualize and categorize the diverse real-world challenges faced by the P-VI/blindness in indoor settings, and the main functional and non-functional requirements to be considered for implementing an assistive solution for indoor scene understanding. We also proposed a design framework with a reference architecture (Fig. 1–Solution Architecture), that can help researchers in this field to understand the necessary components and tasks, and the various approaches that could be used in building a scene understanding solution for P-VI/blindness to better understand their surroundings. Then some guidelines for the selection of appropriate technologies for the development of such solutions were discussed which consider different aspects such as the effectiveness of the algorithms and devices, their limitations, computational load, and estimated cost for the hardware needed for the implementation. Afterwards, we applied the proposed framework to an existing solution as a showcase of how this method could be applied for the assessment of and comparison between assistive solutions.

We additionally discussed some other factors that can affect the utility of these solutions. Among the most important factors are generalizability and user experience. During the past few years many solutions have been offered by various researchers to the community who are blind. However, the current solutions mainly work in limited environments and for limited use cases. For instance, the performance of the methods for object detection or depth detection may vary in different environments.

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. This will provide valuable insights into areas for improvement of the framework to better serve its intended users.

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.

Author contributions Authorship for this manuscript follows the principles and authorship practices outlined by Springer Nature. All authors contributed to the intellectual development of the work, provided essential feedback, and approved the final version for publication. Mohammad Moeen Valipoor wrote the main manuscript text. Angelica de Antonio and Julián Cabrera reviewed and revised the paper.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Open Access Funding provided by Universidad Politécnica de Madrid through an open access agreement with Springer Nature. This work was partially supported by Project PID2020-115132RB (“SARAOS”) funded by MCIN/AEI/10.13039/501100011033 of the Spanish Government.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical approval The user research was granted approval by the Ethics Committee of Universidad Politécnica de Madrid under reference number 2022–077. Each participant approved a consent form in order to participate in the interviews.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Pascolini, D., Mariotti, S.P.: Global estimates of visual impairment: 2010. *Br. J. Ophthalmol.* **96**, 614–618 (2012). <https://doi.org/10.1136/bjophthalmol-2011-300539>
- Nguyen, M., Le, H., Yan, W.Q., Dawda, A.: 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*, 1, 8–13 (2019). <https://doi.org/10.1109/M2VIP.2018.8600857>
- Valipoor, M.M., de Antonio, A.: Recent trends in computer vision-driven scene understanding for VI/blind users: a systematic mapping. *Univers. Access. Inf. Soc.* (2022). <https://doi.org/10.1007/s10209-022-00868-w>
- Hersh, M.A., Johnson, M.A.: 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.* **7**, 252609 (2010). <https://doi.org/10.1080/11762322.2010.523626>
- Bhowmick, A., Hazarika, S.M.: An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends. *J. Multimodal. User. Interface.* **11**, 149–172 (2017). <https://doi.org/10.1007/s12193-016-0235-6>
- Conradie, P., Mioch, T., Saldien, J.: Blind user requirements to support tactile mobility. *CEUR. Workshop. Proc.* (2014). <https://doi.org/10.13140/2.1.2492.3845>
- Wang, S., Yu, J.: Everyday information behaviour of the visually impaired in China. *Inf. Res.* **22**(1), 743 (2017). Retrieved from <http://InformationR.net/ir/22-1/paper743.html> (Archived by WebCite® at <http://www.webcitation.org/6pFtXbqJr>)
- Liu, Y., Stiles, N.R.B., Meister, M.: Augmented reality powers a cognitive assistant for the blind. *Elife* **7**, e37841 (2018). <https://doi.org/10.7554/eLife.37841>
- Mekhalfi, M.L., Melgani, F., Bazi, Y., Alajlan, N.: Fast indoor scene description for blind people with multiresolution random projections. *J. Vis. Commun. Image Represent. Commun. Image. Represent.* **44**, 95–105 (2017). <https://doi.org/10.1016/j.jvcir.2017.01.025>
- Khairnar, DP., Karad, RB., Kapse A, et al.: 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 pp. 32–37. Pune, India (2020)
- Imtiaz, MA., Aziz, S., Zaib, A., et al.: Wearable scene classification system for visually impaired individuals 2nd international conference on electrical communication and computer engineering. ICECCE. (2020). <https://doi.org/10.1109/ICECCE49384.2020.9179439>
- Presti, G., Ahmetovic, D., Ducci, M., et al.: Watchout: obstacle sonification for people with visual impairment or blindness. *ASSETS 2019—21st International ACM SIGACCESS Conference on Computers and Accessibility* 402–413 (2019). <https://doi.org/10.1145/3308561.3353779>
- Apple unveils ARKit 2. <https://www.apple.com/newsroom/2018/06/apple-unveils-arkit-2/> (2007). Accessed 2 Mar 2024
- Sarwar, M.G., Dey, A., Das, A.: Developing a LBPH-based face recognition system for visually impaired people 2021 1st International Conference on Artificial Intelligence and Data Analytics, CAIDA 2021 286–289 (2021). <https://doi.org/10.1109/CAIDA51941.2021.9425275>
- Chen, Z., Liu, X., Kojima, M., et al.: A wearable navigation device for visually impaired people based on the real-time semantic visual slam system. *Sensors* **21**, 1–14 (2021). <https://doi.org/10.3390/s21041536>
- Abraham, L., Mathew, N. S., George L, Sajan, S. S. 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, pp 972–976. India (2020)
- Envision. <https://www.letsenvision.com/> (2023). Accessed 14 Dec 2023
- Seeing AI. <https://www.microsoft.com/en-us/ai/seeing-ai/> (2023). Accessed 14 Dec 2023
- Lookout—Assisted vision. <https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en&gl=US> (2023). Accessed 13 Dec 2023
- Aira. <https://aira.io/> (2023). Accessed 14 Dec 2023
- Be My Eyes. <https://www.bemyeyes.com/> (2023). Accessed 14 Dec 2023
- Stangl, A., Shiroma, K., Davis, N., et al.: Privacy concerns for visual assistance technologies. *ACM. Trans. Access. Comput.* (2022). <https://doi.org/10.1145/3517384>
- Norlund, T., Hagström, L., Johansson, R.: 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 149–162 (2021). <https://doi.org/10.18653/V1/2021.BLACKBOXNLP-1.10>
- Wise, E., Li, B., Gallagher, T., 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, pp 1–7 (2012).
- Hersh, M.A., Johnson, M.A.: 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.* **7**, 277–288 (2010). <https://doi.org/10.1080/11762322.2010.523626>
- Ruffieux, S., Hwang, C., Junod, V., et al.: 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.* **22**, 463–475 (2023). <https://doi.org/10.1007/s10209-021-00857-5>
- Akter, T.: Privacy considerations of the visually impaired with camera based assistive tools. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* 69–74 (2020). <https://doi.org/10.1145/3406865.3418382>
- Jeamwathanachai, W., Wald, M., Wills, G.: Indoor navigation by blind people: behaviors and challenges in unfamiliar spaces and buildings. *Br. J. Vis. Impair.* (2018). <https://doi.org/10.1177/0264619619833723>
- Alamri, A.: 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) 38–42 (2023)
30. Szpiro, S., Zhao, Y., Azenkot, S.: 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. Association for Computing Machinery, pp 61–72, New York (2016)
 31. Tong, K., Wu, Y.: Deep learning-based detection from the perspective of small or tiny objects: a survey. *Image Vis. Comput. Comput.* **123**, 104471 (2022). <https://doi.org/10.1016/j.imavis.2022.104471>
 32. Silva, C., Wimalaratne, P.: Context-aware assistive indoor navigation of visually impaired persons. *Sens. Mater.* **32**, 1497 (2020). <https://doi.org/10.18494/SAM.2020.2646>
 33. Mashiat, M., Ali, T., Das, P., et al.: Towards assisting visually impaired individuals: a review on current status and future prospects. *Biosens. Bioelectron. X* **12**, 100265 (2022). <https://doi.org/10.1016/j.biosx.2022.100265>
 34. Ntakolia, C., Dimas, G., Iakovidis, D.K.: User-centered system design for assisted navigation of visually impaired individuals in outdoor cultural environments. *Univers. Access. Inf. Soc.* (2020). <https://doi.org/10.1007/s10209-020-00764-1>
 35. Bajpai, V., Gorthi, R. P.: On non-functional requirements: a survey. 9–12 (2012)
 36. Tapu, R., Mocanu, B., Zaharia, T.: Wearable assistive devices for visually impaired: a state of the art survey. *Pattern. Recognit. Lett.* **137**, 37–52 (2020). <https://doi.org/10.1016/j.patrec.2018.10.031>
 37. Ohn-Bar, E., Kitani, K., Asakawa, C.: Personalized Dynamics Models for Adaptive Assistive Navigation Systems. In: Conference on Robot Learning (2018)
 38. Shen, J., Dong, Z., Qin, D., et al.: *ivision: an assistive system for the blind based on augmented reality and machine learning*. Springer International Publishing (2020)
 39. Khan, A., Khusro, S.: An insight into smartphone-based assistive solutions for visually impaired and blind people: issues, challenges and opportunities. Springer, Berlin Heidelberg (2020)
 40. Shinohara, K., Wobbrock, J. O.: In the shadow of misperception: assistive technology use and social interactions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, pp. 705–714. New York (2011)
 41. dos Santos, A.D.P., Ferrari, A.L.M., Medola, F.O., Sandnes, F.E.: Aesthetics and the perceived stigma of assistive technology for visual impairment. *Disabil. Rehabil. Assist. Technol.* **17**, 152–158 (2022). <https://doi.org/10.1080/17483107.2020.1768308>
 42. Kuriakose, B., Shrestha, R., Sandnes, F. E.: 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). pp 2464–2470 (2021)
 43. Pawar, P. G., Devendran, V.: Scene understanding: a survey to see the world at a single glance. 2019 2nd international conference on intelligent communication and computational techniques, ICCT 2019 182–186. (2019) <https://doi.org/10.1109/ICCT46177.2019.8969051>
 44. Xie, L., Lee, F., Liu, L., et al.: Scene recognition: a comprehensive survey. *Pattern. Recognit.* (2020). <https://doi.org/10.1016/j.patcog.2020.107205>
 45. Liu, Y., Chen, Q., Chen, W., Wassell, I.: 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. AAAI Press (2018)
 46. Shi, J., Zhu, H., Yu, S., et al.: Scene categorization model using deep visually sensitive features. *IEEE Access* **7**, 45230–45239 (2019). <https://doi.org/10.1109/ACCESS.2019.2908448>
 47. Seong, H., Hyun, J., Kim, E.: FOSNet: An end-to-end trainable deep neural network for scene recognition. *IEEE Access* **8**, 82066–82077 (2020). <https://doi.org/10.1109/ACCESS.2020.2989863>
 48. Tan, M., Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning. PMLR, pp. 6105–6114 (2019)
 49. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016)
 50. Kuipers, B.: The cognitive map: could it have been any other way? *Spat. Orientat.* (1983). <https://doi.org/10.1007/978-1-4615-9325-6>
 51. Tan, S., Guo, D., Liu, H., et al.: Embodied scene description. *Auton. Robot. Robot.* **46**, 21–43 (2022). <https://doi.org/10.1007/s10514-021-10014-9>
 52. Delloul, K., Larabi, S.: Egocentric scene description for the blind and visually impaired. In: 2022 5th international symposium on informatics and its applications (ISIA). pp 1–6 (2022)
 53. Hersch, M.: Mental maps and the use of sensory information by blind and partially sighted people. *ACM. Trans. Access. Comput.* (2020). <https://doi.org/10.1145/3375279>
 54. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). pp 770–778 (2016)
 55. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. *Commun. ACM.* **60**, 84–90 (2017). <https://doi.org/10.1145/3065386>
 56. Zhao, Z.Q., Zheng, P., Xu, S.T., Wu, X.: Object detection with deep learning: a review. *IEEE. Trans. Neural. Netw. Learn. Syst.* **30**, 3212–3232 (2019). <https://doi.org/10.1109/TNNLS.2018.2876865>
 57. Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on computer vision and pattern recognition (CVPR). pp 2818–2826 (2016)
 58. Bhumbla, S., Gupta, D. K., 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., pp 827–832 (2023)
 59. He, K., Gkioxari, G., Dollar, P., Girshick, R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
 60. Venkataramanan, A.K., Facktor, M., Gupta, P., Bovik, A.C.: Assessing the impact of image quality on object-detection algorithms. *Electron. Imaging.* **34**, 334–1–334-1 (2022). <https://doi.org/10.2352/EI.2022.34.9.IQSP-334>
 61. Guo, H., Lu, T., Wu, Y. Dynamic low-light image enhancement for object detection via end-to-end training in: 2020 25th international conference on pattern recognition (ICPR), pp. 5611–5618 (2021)
 62. Hafiz, A.M., Bhat, G.M.: A survey on instance segmentation: state of the art. *Int. J. Multimed. Inf. Retr.* **9**, 171–189 (2020). <https://doi.org/10.1007/s13735-020-00195-x>
 63. Kompalli, P.L., Kalidindi, A., Chilukala, J., et al.: A color guide for color blind people using image processing and openCV. *ijOE* **19**, 30–46 (2023). <https://doi.org/10.3991/ijoe.v19i09.39177>
 64. Allam, M., ElShaarawy, I., Farghal, S.A.: In: Magdi Dalia, A., Helmy, Y.K., M. M., J. A. (eds.) Digital Transformation Technology, pp. 195–216. Springer, Singapore (2022)

65. Widayani, A., Kusuma, H., Purwanto, D.: Visually impaired person detection using deep learning for dangerous area warning system. In: 2022 international seminar on intelligent technology and its applications: advanced innovations of electrical systems for humanity, ISITIA 2022—Proceeding. institute of electrical and electronics engineers Inc., pp 204–208 (2022)
66. Chung, M. A., Chai, S. Y., Hsieh, M. C. et al.: Road Pothole Detection Algorithm and Guide Belt Designed for Visually Impaired. In: 2023 IEEE 3rd International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB). pp. 475–478 (2023)
67. Google cloud vision
68. Microsoft azure computer vision
69. Amazon Rekognition. <https://aws.amazon.com/rekognition/> (2024). Accessed 5 Feb 2024
70. Miles, F. A.: 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 **19**: 407–408 (1996)
71. Scharstein, D., Szeliski, R., Zabih, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In: proceedings IEEE workshop on stereo and multi-baseline vision (SMBV 2001). pp. 131–140 (2001)
72. Adil, E., Mikou, M., Mouhsen, A.: A novel algorithm for distance measurement using stereo camera. CAAI. Trans. Intell. Technol. **7**, 177–186 (2022). <https://doi.org/10.1049/cit2.12098>
73. Cámara Intel® RealSense™ SR300. <https://www.intel.la/content/www/xl/es/products/sku/92329/intel-realsense-camera-sr300/specifications.html> (2024). Accessed 8 Feb 2024
74. Maculotti, G., Ulrich, L., Olivetti, E.C., et al.: A methodology for task-specific metrological characterization of low-cost 3D camera for face analysis. Measurement. (Lond.) (2022). <https://doi.org/10.1016/j.measurement.2022.111643>
75. A Brief Analysis of the Principles of Depth Cameras: Structured Light, TOF, and Stereo Vision. https://wiki.dfrobot.com/brief_analysis_of_camera_principles (2024). Accessed 25 Feb 2024
76. Azure kinect depth camera
77. Azure Kinect DK hardware specifications. <https://learn.microsoft.com/en-us/azure/kinect-dk/hardware-specification> (2024). Accessed 21 Jan 2024
78. Intel RealSense LiDAR Camera L515. <https://intelrealsense.com/lidar-camera-l515/> (2024). Accessed 21 Jan 2024
79. Khan, F., Salahuddin, S., Javidnia, H.: Deep learning-based monocular depth estimation methods—A state-of-the-art review. Sensors (2020). <https://doi.org/10.3390/s20082272>
80. Khan, F., Salahuddin, S., Javidnia, H.: Deep learning-based monocular depth estimation methods—A state-of-the-art review. Sensors (Switzerland) **20**, 1–16 (2020). <https://doi.org/10.3390/s20082272>
81. Unsupervised monocular depth estimation in highly complex environments. <https://en.x-mol.com/paper/article/1420832096734703616> (2022). Accessed 29 Nov 2022
82. Ming, Y., Meng, X., Fan, C., Yu, H.: Deep learning for monocular depth estimation: a review. Neurocomputing **438**, 14–33 (2021). <https://doi.org/10.1016/J.NEUCOM.2020.12.089>
83. Lin, B.S., Lee, C.C., Chiang, P.Y.: Simple smartphone-based guiding system for visually impaired people. Sensors (Switzerland) (2017). <https://doi.org/10.3390/s17061371>
84. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: MonoSLAM: Real-time single camera SLAM. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 1052–1067 (2007). <https://doi.org/10.1109/TPAMI.2007.1049>
85. Kuriakose, B., Shrestha, R., Sandnes, F.E.: In: Antona, M., Stephanidis, C. (eds.) LiDAR-Based Obstacle Detection and Distance Estimation in Navigation Assistance for Visually Impaired, pp. 479–491. Springer International Publishing, Cham (2022)
86. Hakim, H. Fadhil, A. (2019) Navigation system for visually impaired people based on RGB-D camera and ultrasonic sensor. ACM International Conference Proceeding Series 172–177. <https://doi.org/10.1145/3321289.3321303>
87. Xu, P., Kennedy, G.A., Zhao, F.Y., et al.: Wearable obstacle avoidance electronic travel aids for blind and visually impaired individuals: a systematic review. IEEE Access **11**, 66587–66613 (2023). <https://doi.org/10.1109/ACCESS.2023.3285396>
88. F. A., NADA, A., A. M., MASHALI, S.: Effective fast response smart stick for blind people. Institute of research engineers and doctors, LLC, pp. 5–11 (2015)
89. Khan, S., Nazir, S., Khan, H.U.: Analysis of navigation assistants for blind and visually impaired people: a systematic review. IEEE Access **9**, 26712–26734 (2021). <https://doi.org/10.1109/ACCESS.2021.3052415>
90. Vanitha, M., Rajiv, A., Elangovan, K., Kumar, S.V.: A smart walking stick for visually impaired using raspberry pi. Int. J. Appl. Math. **119**, 3485–3489 (2018)
91. Jin, S.G., Ahmed, M.U., Kim, J.W., et al.: Combining obstacle avoidance and visual simultaneous localization and mapping for indoor navigation. Symmetry (Basel) **12**, 1–13 (2020). <https://doi.org/10.3390/SYM12010119>
92. Jia, Y., Yan, X., Xu, Y.: A Survey of simultaneous localization and mapping for robot. In: 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). pp. 857–861 (2019)
93. Atoui, O., Husni, H., Mat, R.C.: Visual-based semantic simultaneous localization and mapping for Robotic applications a review. AIP Conf. Proc. (2019). <https://doi.org/10.1063/1.5121082>
94. Rui, C., Liu, Y., Shen, J., et al.: 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 409–414 (2021). <https://doi.org/10.1109/PIC53636.2021.9687018>
95. Campos, C., Elvira, R., Rodriguez, J.J.G., et al.: ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM. IEEE Trans. Robot. **37**, 1874–1890 (2021). <https://doi.org/10.1109/TRO.2021.3075644>
96. Raj, T., Hashim, F.H., Huddin, A.B., et al.: A survey on LiDAR scanning mechanisms. Electronics (Basel) (2020). <https://doi.org/10.3390/electronics9050741>
97. Google Tesseract An optical character recognition (OCR) engine (2015)
98. Anwar, N., Khan, T., Mollah, A.F.: Text Detection from Scene and Born Images: How Good is Tesseract? In: Pundir, A.K.S., Yadav, N., Sharma, H., Das, S. (eds.) Recent Trends in Communication and Intelligent Systems, pp. 115–122. Springer Nature, Singapore (2022)
99. Neat, L., Peng, R., Qin, S., Manduchi, R.: Scene text access: a comparison of mobile ocr modalities for blind users. In: proceedings of the 24th international conference on intelligent user interfaces. association for computing machinery, New York, pp. 197–207 (2019)
100. Madake, J., Bhatlawande, S., Solanke, A., Shilaskar, S.: A qualitative and quantitative analysis of research in mobility technologies for visually impaired people. IEEE. Access. **11**, 82496–82520 (2023). <https://doi.org/10.1109/ACCESS.2023.3291074>
101. Elmannai, W., Elleithy, K.: Sensor-based assistive devices for visually-impaired people: current status, challenges, and future directions. Sensors (2017). <https://doi.org/10.3390/s17030565>
102. Granquist, C., Sun, S.Y., Montezuma, S.R., et al.: Evaluation and comparison of Artificial Intelligence vision aids: Orcam myeye 1 and seeing AI. J. Vis. Impair. Blind. **115**, 277–285 (2021). <https://doi.org/10.1177/0145482X211027492>

103. Lee, J., Herskovitz, J., Peng, YH., Guo, A.: 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 (2022)
104. Xie, L., Lee, F., Liu, L., et al.: Hierarchical coding of convolutional features for scene recognition. *IEEE. Trans. Multimed.* **22**, 1182–1192 (2020). <https://doi.org/10.1109/TMM.2019.2942478>
105. Wang, C. Y., Bochkovskiy, A., Liao, H. YM.: YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. *Institute of Electrical and Electronics Engineers (IEEE)*, pp. 7464–7475 (2023)
106. YoloV7-ncnn-Raspberry-Pi-4. <https://github.com/Qengineering/YoloV7-ncnn-Raspberry-Pi-4> (2024). Accessed 18 Jan 2024
107. Detectron2 Model Zoo and Baselines. https://github.com/facebookresearch/detectron2/blob/main/MODEL_ZOO.md (2024). Accessed 19 Jan 2024
108. Caesar, H., Uijlings, J., Ferrari, V.: COCO-Stuff: Thing and Stuff Classes in Context. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1209–1218 (2018)
109. Cakic, S., Popovic, T., Krco, S., et al.: Developing edge AI computer vision for smart poultry farms using deep learning and HPC. *Sensors* (2023). <https://doi.org/10.3390/s23063002>
110. Cabanillas-Carbonell, M., Chávez, A. A., Barrientos, J. B.: 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). pp. 1–5 (2020)
111. Intel® RealSense™ Depth Camera D455. <https://www.intelrealsense.com/depth-camera-d455/>. Accessed 19 Jan 2024
112. Ranftl, R., Lasinger, K., Hafner, D., et al.: Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 1623–1637 (2022). <https://doi.org/10.1109/TPAMI.2020.3019967>
113. MiDAS Github Repository. <https://github.com/isl-org/MiDaS>. Accessed 21 Jan 2024
114. Beshley, M., Volodymyr, P., Beshley, H., Gregus, M.: A smart-phone-based computer vision assistance system with neural network depth estimation for the visually impaired. pp. 26–36 (2023)
115. Saranya, S., Sudha, G., Subbiah, S.: Raspberry Pi based smart walking stick for visually impaired person. *AIP Conf. Proc.* **2520**, 020010 (2022). <https://doi.org/10.1063/5.0103097>
116. Kunta, V., Tuniki, C., Sairam, U.: Multi-functional blind stick for visually impaired people. In: 2020 5th international conference on communication and electronics systems (ICCES). pp. 895–899 (2020)
117. Intel® RealSense™ Depth Camera D415
118. Hussain, S. S., Durrani, D., Khan, A. A., et al.: In-door obstacle detection and avoidance system for visually impaired people. In: 2020 IEEE Global Humanitarian Technology Conference (GHTC). pp. 1–7 (2020)
119. Rajesh, M., Rajan, B. K., Roy, A., 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). pp. 1–5 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.