# Image Recognition Tools for Blind and Visually Impaired Users: An Emphasis on the Design Considerations.

Sandra Fernando*
Assistive Technology Research Group, SCDM, London Metropolitan University, s.fernando@londonmet.ac.uk

Chiemela Ndukwe
Assistive Technology Research Group, SCDM, London Metropolitan University, c.ndukwe1@londonmet.ac.uk

Bal Virdee
Assistive Technology Research Group, SCDM, London Metropolitan University, b.virdee@londonmet.ac.uk

Ramzi Djemai
SCDM, London Metropolitan University, r.djemai@londonmet.ac.uk

The existing digital image recognition technologies available for blind individuals are commercially accessible but still at an immature stage, necessitating enhancements in their capabilities. Areas requiring improvement in current image recognition tools for the visually impaired encompass information accuracy, information adequacy, appropriateness of information for blind individuals, functional sufficiency, and ergonomic suitability. This research endeavors to explore technology employing an inclusive approach to overcome limitations inherent in current digital image recognition technologies for the visually impaired. To streamline the research process, the initial phase involves a critical evaluation of deficiencies present in existing image recognition technologies for blind/visually impaired users through primary and secondary investigations. This strategic evaluation aims to identify key deficiencies, guiding assistive technology developers in focusing their efforts. Simultaneously, a survey is conducted to establish a comprehensive checklist of usability features and requirements for the proposed technology, aligning with the ISO 9241-110 standard. The overarching goal of this research is to address the question, "What design considerations should be taken into account in designing image recognition technology for blind and visually impaired people?" The research outcomes are intended to establish a standard for designers of digital products catering to blind/visually impaired users, fostering improved awareness and shaping attitudes toward individuals with visual disabilities in the development of image recognition software.

**CCS CONCEPTS** • Human-centred computing • Accessibility

**Additional Keywords and Phrases:** Image recognition tool, visually impaired, Blind users, Assistive technology

**ACM Reference Format:**

Sandra Fernando, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock

* the footnote text for the author (if applicable) here.

’18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

# 1 Introduction

Digital images have become a major part of digital communication and part of daily life nowadays. However, digital images could be considered a major barrier to individuals who have visual impairments because they cannot access information presented in digital images. Technological development in the field of image recognition has grown considerably recently, and much of the reason is that it can quickly categorise and identify information in digital images. It is of great importance that the design considerations for image recognition technology for people who are blind and visually impaired form part of the broader problem of inclusivity facing the digitally disadvantaged.

Image recognition systems can reduce a communication gap between people with and without vision by translating any visual content in digital media to formats friendly to the blind. Still, the usefulness of digital image recognition software for the totally blind or people with low vision is limited. The power of artificial intelligence (AI) and algorithms in recognising and communicating complex information in pictures, like abstract or metaphorical images, can be relatively low [1]. The overall performance of the image recognition feature depends largely on the quality of the image itself, especially in terms of clarity and lighting conditions.

Ongoing investigations hold promise in facilitating the development of digital platforms designed to inclusively engage with digital imagery, specifically tailored to meet the needs of individuals with visual impairments.

Recent works focusing on the application of AI, such as computer vision for people who are blind or have low vision (BLV), show a weak fit between the technological focus and the actual needs of users who are BLV [2]. An analysis of 646 papers and interviews with 24 BLV participants reveals preferences towards AI that uses conversational interfaces and head-mounted devices, hence pointing at research gaps in focus [2].

The current research into persons with visual disabilities aids in the structuring of key topics and factors [3] [4], such as: (1) conceptual frameworks centred on the idea that people can access and use digital technology with digital media and the internet [5]. There are two dimensions of the importance of digital inclusion: (1) access to some tools and services is seen as a must and allows one to be an involved member of society these days [6] [7]; and (2) visual disability. As one more method to consider the overall concept of cognitive processes, visual disability represents a wide spectrum of deficits having different negative implications, namely an absolute lack of light perception, colour vision disorders, and general low acuity. In short, the inability of individuals with visual impairments to access visual information on digital platforms is a big barrier to digital inclusion. These research activities collectively contribute to advancing understanding and addressing the multifaceted challenges associated with digital inclusion for individuals with visual impairments.

Understanding images for individuals with visual impairments is one of the key aspects of the development of assistive technologies. In order to understand the potential benefits and limitations of image recognition systems for the visually impaired, one has to consider the various modes in which a visually impaired person sees and interprets visual information.

The International Organisation for Standardisation (ISO) has been very instrumental in developing a wide range of standards on assistive technologies and image analysis tools specifically designed for use by the visually impaired population [8] [9]. Such standards make it possible to have assistive technologies that are flexible, inclusive, and accommodating to the varied perception modes of the large population of visually impaired people.

This availability of information in a visual form to a visually impaired person can be made possible with alternative display technologies such as braille displays or audio output. Besides, ISO standard 9241-171 places a lot of emphasis on tailoring assistive technology specifically to the unique needs and preferences of the individual user. This would be achieved through the application of user-centered design principles and working with key users who have vision impairments. ISO standards help guide in creating assistive technology that would be of aid

to visually impaired people in, among other things, the compatibility with various modes of perception and being adapted to individual needs.

The most recent working from OpenAI is the GPT-4V (Vision) system, which is one of the unique ones among its predecessors, where GPT-4 users can instruct to give directions about textual context for how the provided image inputs by the said user should be analyzed [10]. Within the understanding of where to go with the given images, given another textual context, this is a hallmark by itself of A.I [10]. Building on the foundational work done for GPT-4, extensive evaluations of the capacity of GPT-4 with respect to safety properties are done [10]. Focus here is more on image inputs: robustness and reliability [10]. Appropriately, one research done by Gamage et al. went further in establishing the actual real needs of the blind and low-vision assistance user for smart devices [11]. Using at least 646 recent studies, the authors looked into tasks to everything from object detection to devices that would best be suitable for this user category [2]. Interestingly, these findings demonstrated that the preference was toward conversational agent interfaces and head-mounted devices [2]. Research of this kind is what bridged this gap between technological advancements and the practical needs of the visually-handicapped so that their daily experience could be improved [2].

Modalities are there to support various cognitive processing styles and needs in developing image recognition systems [12]. Concretely, Visual Modalities: Image recognition systems should use the visual modalities to allow for object recognition, classification, and segmentation tasks for the information being processed. The designer may consider presenting information in the form of an image or video intended for partially sighted people. But this should not be the only way to provide information when creating a system for blind users. Text Modalities: A textual modality can take text input, which could be anything from captions, descriptions, or metadata that are related to images. Image recognition systems that incorporate textual modalities can help describe or add captions to the images and aid in content understanding as well as semantic retrieval, as some screen readers do [12]. For this, blind people need text-to-speech software. Modalities: This is the best-preferred and most-used output modality for blind users. Image recognition systems that incorporate auditory modalities that enable audible description or feedback to support a user's interaction with visual information with spoken commands, auditory cues, or audio-based interfaces. Multimodal integration means putting together many modalities, that is, visual, textual, and auditory, for cognitive process reinforcement and to cater for a distinctive blend of user needs. However, these are to be integrated as user-preferred optional activities rather than modality overloading. Combining several modalities in recognition systems allows for richer and more integral representation of visual content, catering to a wider range of differences in cognitive processing styles and preferences among users [12]. This enables accessing and interacting with visual information in any modality preference, be it visual, textual, or auditory channels, according to them [12].

Meanwhile, the current project looks at the relationship among the major concepts and variables, investigating the capabilities of the image recognition technology that will help include the visually impaired in the digitised world. It points the way to the chances of identifying barriers to the use of effective and timely technology for image recognition for the visually impaired, considering both accessibility and technological limitations. The present study exposed the participants to the various systems that have been developed to offer help in a bid to respond to the needs of the blind and visually impaired through the recognition of images. Some of the technologies tested on the platforms included AIpoly, Microsoft's Seeing AI, Audible Vision, Supersense, Google's Vision AI, Envision, and Amazon's Amazon Rekognition, among others, which in total summed up to twenty-one different platforms. Participants were not required to test every system; instead, they reviewed tools based on their individual familiarity and prior usage. This approach was chosen to gather authentic feedback on usability and effectiveness from the perspective of regular users who have developed a certain level of proficiency with specific tools. The manner in which the users were enabled to input to these systems ranged from voice commands to touch-based interfaces, in line with the system interface and the diverse ways such technologies catered to the needs of a visually impaired user.

While it is a profound understanding of the interrelations between these concepts, the work tries to present the potential of image recognition technology to enhance the inclusion of the visually impaired in the digital space.

Besides, the study attempts to identify areas that require improvement. This is a systematic process to evaluate the challenges encountered by the visually challenged in using photo recognition technology in an effort to offer invaluable insights into the improvement and advancement of technological solutions designed for the digital inclusion of such people.

The paper is organized into some major segments so that the reach to an analysis of image recognition tools and ergonomic design standards can be effectively made. The introduction sets the stage by outlining the purpose and scope of the study. It also includes the description of image recognition tools, which notwithstanding the current technological state in which they are, address subsections about their strengths and weaknesses, the AI algorithms by which they are underpinned, and the corresponding ISO standards for ergonomic design. Methodology and study design detail how the research had been implemented and what methods it used to collect data in a given procedure. The findings section is followed by a discussion and conclusion recasting the findings with their implications and possible directions for future research.

# 2 OVERVIEW OF IMAGE RECOGNITION TOOLS

These tools are collectively sometimes known as image recognition or assistive technology, and, although there might be differences between the two, in general, they aim for the same purpose of allowing the visually impaired to gain experience of the visual world. of a number of image recognition tools and what features they have. Each tool is assessed for features that are crucial for empowering people with visual impairments, such as voice output, text reading, object recognition, currency identification, environmental description, barcode recognition, colour identification, face analysis, handwriting recognition, and voice commands. An asterisk indicates that a particular feature is present in the tool. This comparison overview brings out lucidly the capabilities of each tool, providing insight into their applicability in different needs and contexts.

**Table 1: Tools and characteristics of commonly used IR tools**

| Name | Voice results | Reads Text | Object detection | Currency recognition | Describes environment | Identify barcodes | Colour detection | Face analysis | Reads handwriting | Voice Commands |
|---|---|---|---|---|---|---|---|---|---|---|
| AIpoly | * | * | * | * | | | * | | | |
| Seeing AI by Microsoft | * | * | * | * | * | * | * | * | * | * |
| Audible vision | * | * | | | | * | | | | |
| Supersense | * | * | | * | * | * | | | | |
| Vision AI by Google | | * | * | * | * | * | * | * | * | |
| Envision | * | * | | | * | | | | * | |
| Amazon Rekognition by Amazon | | * | * | | * | | | * | | |
| BrookesTalk | * | * | | | | | | | | |

| Name | Voice results | Reads Text | Object detection | Currency recognition | Describes environment | Identify barcodes | Colour detection | Face analysis | Reads handwriting | Voice Commands |
|---|---|---|---|---|---|---|---|---|---|---|
| LookTel | * | | * | * | | * | | | | |
| CVAT | | | * | | | | | | | |
| Roboflow | | | * | | | | | | | |
| pwWebSpeak | * | * | | | | | | | | |
| IBM Home Page Reader by IBM | * | * | | | | | | | | |
| WebAnywhere | * | * | | | | | | | | |
| Connect Outloud | * | * | | | | | | | | |
| JAWS | * | * | | | | | | | | |
| TapTapSee | * | | * | | | | * | | | |
| Talking Goggles | * | * | * | | | | * | | | |
| Be My Eyes | * | * | | | * | | | | | |
| Sullivan+ | | * | | | * | | * | * | | |
| VizWiz | * | * | * | * | * | * | * | * | * | * |

The above Table 1 does not only include the tools that blind people use but also contains lists of tools for voice web browsers like IBM Home Page Reader, WebAnywhere, BrookesTalk, pwWebSpeak, Connect Outloud, and image annotation tools, to name just a few, comprising CVAT, Roboflow, and others in the least. JAWS has no other features beyond voice results and reading the text as presented in Table 1, but has the capability of describing images and OCR, which did not explicitly fall among the identified primary features. In evaluating the analysis in Table 1 above and considering the different image recognition tools invented to assist the visually impaired, some of the features emerge as both most widespread and critical. Predominantly, the functionalities 'Voice results' and 'Reads Text' are featured across most of the analysed platforms, which strongly indicates this as a kind of basic feature in converting visual data into information comprehensible for the sound center. This highlights the general importance of auditory feedback in tools for users with visual impairments. On the other hand, more specialised features such as 'Describes environment' and 'Face analysis' are not as common, even though they are expected to provide users with a greater amount of more in-depth contextual information. In fact, the existence of a gap regarding the choices available in the market today, as reflected in the increased – compared to the basic tools – requirements of users in tools' features, has been made clear; under these terms, the selection of technology that will serve the needs of visually impaired users should be based on strict criteria. In this step, the choice of specific image recognition tools for blind users becomes a crucial one.

This choice is not simply a matter of preference but rather a crucial element in promoting access and independence for persons with visual impairments. Therefore, incorporating these more advanced, though not so widely applicable, features in existing tools would lead to a greater impact of use and address the specific needs of this group. Some of the issues that guided the choice of the tools are indicated among the many that exist.

Features: Its lightweight design and real-time recognition make it fit to be fully included in the lives of users and also embedded in their mobile devices. Seeing AI by Microsoft: One of the largest technology brands, this epitomises an inclusivity-based app, with features for text reading, scene description, and even currency recognition. This is made more effective through its use as part of the Microsoft ecosystem. Audible Vision: It directly caters to blind users using its voice results and precise way of reading text. The application that perhaps best illustrates the need for practical development is its use when combined with colour detection in Supersense. Google Vision AI: Works by a strong model in machine learning to give its best object detection, and in a good level of recognizing text in accordance with the industry norm. Relating: Adjustability to multi-language with consistent updates: Envision should remain relevant to the user with time, fulfilling emerging needs. Usability and User Experience: Amazon Rekognition by Amazon: While not a tool specifically built for the visually impaired, it can still define objects. Additionally, Amazon Rekognition is an insightful cloud-based service managed on a no-end insight foundation. Such integrations, working alongside Braille displays and screen readers, will help complement communication and improve the ease of use for users with visual impairments. LookTel: The user-friendly interface and efficient object recognition contribute to a positive user experience. JAWS: Is a screen reader that is high quality, versatile among various applications and settings, and therefore can be tailored to individual users. Community Support and Adoption: Be My Eyes: This is a worldwide network of volunteers who form their users into sighted helpers for everyday tasks and activities that many people generally take for granted in their everyday lives. In March 2023, Be My Eyes and OpenAI launched "Be My AI," integrating GPT-4V with the Be My Eyes platform to describe visuals for the blind and low-vision community. Testing through August had shown engagement from 16,000 users [10]. It was, however, during this potential beta test that AI errors and limitations were shown. This leads to continual improvement in the output, with warnings not to use the AI for critical safety tasks and exploring ways to provide responsible descriptions of people that do not compromise privacy [10]. VizWiz, the answer will likely come from what crowdsourcing and text recognition will be able to empower the user to do. Accessibility and Customisation: Sullivan+: An app that is supposed to do a lot of things, including recognising texts and analysing features and faces with the help of money, people, and AI. TapTapSee: Its simplicity—take a photo, get an answer—makes it quick to decide—something important to blind users. Some of the selective features of these image recognition tools include, but are not limited to, their remarkable accuracy, ease of use, and being community-driven, all of which combine to assure the visually impaired of an exceptionally improved experience. Such features make people with visual challenges understand better that technological advancements are making it possible for them to interact with the world in a better, improved way with correct image recognition.

Following this, it becomes apparent that there are important design considerations based on available tools of which the implementation of artificial intelligence (AI) algorithms becomes apparent: The choice of AI algorithms is paramount because they play a crucial role in the performance and usability of image recognition systems [13]. Thus, the integration of advanced AI algorithms is not just a technical decision but a strategic one that directly impacts the quality of life for users reliant on these technologies. Accurate real-time processing come handy for the algorithms good at running for the blind users. The designers will have to strive to use the smallest efficient model, which will mean getting models that run well on mobile devices yet still maintain accuracy. This then implies understanding the trade-off between different algorithms guiding decisions on the design level. For instance, very high accuracy results from a deep learning model may call for very expensive computations and hence unaffordable and unresponsive for runtime execution [14]. Preprocessing Techniques: It includes noise reduction, contrast adjustment, and enhancement of images [15]. Many types of preprocessing improve different features, for example, text and objects for the blind user, and at the same time reduce irrelevant details [15]. Relevance: Preprocessing stages should be such that it ensures the enrichment to the real information to the blind users. For example, the same might be subjected to glow highlights, boundary enhancements to the objects, and other such operations. It is recommended that designers should, hence, try the hybrid feature representations, along with the CNN-based representations using the domain-specific cues of the spatial layout or the semantic context [16]. Such integration strategies then do robust recognition for diverse scenarios. Feature Extraction and

Representation: The architecture then extracts meaningful features from images [17]. Convolutional neural networks associate with a good result to learn features in a hierarchy [17]. Besides that, the blind user should also get related ones: the spatial context, with features, is his object semantics and scene descriptions. Model Explainability: Hence, system feedback is even more relevant for the blind. Best if AI models and systems could throw light wherever possible. Similar methods like the attention map, saliency visualization, and feature contribution come in very handy while trying to understand model decisions. Adaptive Learning and Personalization: The needs for blind users are just too different. Learning systems such as these enables much to be done in allowing the customization of what objects a recognition model learns based on individual tastes, language proficiency, and specific task needs (e.g., reading menus and identifying faces) [18]. Human-AI Collaboration: The other symbiotic benefit is on the blind. In other words, the coupled AI assistance, along with consensus from a human perspective through things like crowdsourcing and volunteer networks, provide for real-time context together with real-time social interaction [19].

## 2.1 Strength of Image Recognition Tools

Image recognition applications come in many forms, and, to a large extent, they aid the visually impaired in recognising various types of objects. Some of the top tools for image identification are powered by Google's Vision AI, which makes it possible for objects to be recognised instantly and accurately. This is one of the greatest and most important aids for blind, low-vision, and colour-blind people to better understand what is in front of them by recognising the objects and colours they meet [20]. This can vary from visors, electronic glasses, and clip-on cameras, which are integrated into numerous combinations, either attaching to spectacles or a head strap. Some of the visor-style systems can help a person with peripheral vision loss gain more of the world, but they are not for on-the-move use. It is advised that these devices be removed when on the move and only be used when stationary, when driving, or when being carried by someone else in a vehicle. Even with these limitations, the devices prove very useful in offering information regarding a person's surroundings at a near distance and are quite important when it comes to the ability to make decisions concerning movement within a room or outside scene [21].

Most of these apps also support money reader functionality, which is able to instantly identify money as well as read out its value. For the identification and counting of money, mainly bills for the blind and the visually impaired, this functionality is enhanced and speedy. When the iPhone's camera is put on a bill, the app quickly speaks out the denomination of the money [22]. Proof of that lies in their ability to identify the currency at once, announce the denomination, and scan barcodes, as has been witnessed [23] [24] [25] [26]. Image recognition APIs can be integrated into most of these tools, which in turn use the camera of the device and make optimum use of voiceover functions. The system generates an image or a video and speaks out the content [27].

## 2.2 Weaknesses of Image Recognition Tools

Some of the drawbacks of the current image recognition technologies are that the tools need to be upgraded. For example, quite a few tools recognise the languages and currencies of only a few countries [28]. The inability of quite a lot of software in the image recognition category to be used with voice commands is a very big drawback, and this restricts its usage by users who are visually impaired [29]. Upgrading is required for object detection, text recognition, and speech output features [30].

Application Form: The image recognition application requires some initial input and activity from the user to work well and obtain the desired results. For instance, the user may have to go through some sort of initial training to differentiate between currencies in different forms. Though this can be a drawback, the application has to be continually configured and tested. To utilize specific programs on an iOS device, users must download them, and prior to commencing item recognition, it is vital for users to establish an image library. This process often involves the assistance of an individual with a keen interest in photography, such as a friend or family member, to curate a repository for housing these photos. It is important to note that this undertaking may require a

substantial investment of time [31]. Another limitation is the time factor regarding image processing after its scanning. The time an image will take before processing after scanning can vary, but is usually between 15 seconds and one minute [32]. Moreover, some wearable assistive technologies come at a very high price, such as £2,695.00 [33]. Other features that the current system lacks include the provision for the ability to interpret any graphical information that may be contained on a web page, for example, a signature. The ability of a system to function without reliance on the internet connection and to respond quickly by performing tasks with few operations. A multi-lingual tool will be a useful interactive enhancement for the user.
In the following section, the AI algorithms used in the image recognition tools (IRT) will be explained.

## 2.3 AI Algorithms Used for Image Recognition

Image recognition involves the use of artificial intelligence techniques that serve a variety of purposes, with specific methodologies applied to specific tasks. OCR is utilised in some recognition technologies to interpret text, and CNNs are used to detect and understand both images and text [34]. The deep neural network-based Image Recognition Technology (IRT) system is designed for the recognition of objects and faces. In this work, a model will be built based on a variant of the ResNet-50 architecture, pre-trained on over three million images from Microsoft (n.d.). ResNet-50 is a variant of Residual Networks, where they have skip connections across some layers. ResNet-50 is quite a complex model that features 50 layers, but it has been successful in differentiating between various objects in images [35] presented residual networks as a novel architectural paradigm to introduce "skip connections" or "shortcuts" into the supposed paths of deep neural networks. Activations can then proceed to pass one or more layers of the network and then skip the other [35]. Residual connections were one of the ideas behind this architectural decision. The ability to gain performance must be maintained at increased network depth, and such residual connections are important in training very deep networks effectively since they alleviate vanishing gradient problems [35]. Image Recognition Technology (IRT) employs very advanced algorithms that utilise natural language processing techniques in formulating text summaries and scene descriptions [36]. The main techniques relied upon are those involving recurrent neural networks and long short-term memory networks that have already been very successful for natural language processing in cases of language translation and sentiment analysis [37] [38].

Image recognition involves the use of artificial intelligence techniques that serve a variety of purposes, with specific methodologies applied to specific tasks. OCR is utilised in some recognition technologies to interpret text, and CNNs are used to detect and understand both images and text [34]. The deep neural network-based Image Recognition Technology (IRT) system is designed for the recognition of objects and faces. In this work, a model will be built based on a variant of the ResNet-50 architecture, pre-trained on over three million images from [34].

The application merges computer vision and natural language processing techniques for detailed and acoustically accurate descriptions of the environment around the user [37]. This function may be particularly useful for people who encounter problems in navigation and perceiving their environment [39]. The contributions given to the IRT area by these high technologies are accessibility and, therefore, inclusiveness for the visually impaired. The MobileNet-SSD combines the lightweight architecture of MobileNet with the single-shot recognition system of SSD. This computational model is targeted to show capable object recognition while running under the stringent conditions typical of mobile and embedded devices. Replacing the base network in the original SSD with MobileNet allows MobileNet-SSD to gain better efficiency than SSD with few compromises in object detection performance. MobileNet integrated with SSDs enables effective real-time object detection even for devices with constrained computational capacities, making it possible due to the model's small size. Real-time object identification is required in almost all robotic, autonomous vehicle, and mobile application domains in which computational resources are very minimal. The MobileNet-SSD model has shown great compatibility in the field of object detection with mobile and embedded systems. The proposed approach integrates a light architecture design of MobileNet with an effective single-shot detection framework proven in prior research to guarantee real-time operation without degradation of accuracy [40] [41]. The MobileNet-SSD combines the energy-efficient

MobileNet architecture with the Single Shot multi-box detection technique, thus forming a hyper-optimised object-detection system that is particularly tailored for real-time performance on mobile devices [40] [41].

An object detection method with the You Only Look Once (YOLO) approach by Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi is one of the most revolutionary innovations in computer vision. It has been known to be fast, accurate, and pioneering for an alternative method to two-stage state-of-the-art methods like R-CNN and Faster R-CNN, which just came into the limelight using the YOLO acronym [42]. Importantly, among the more popular YOLO algorithms, it is highly computational and provides precision; it is best used in applications dealing with real-time work and the need for high processing [43]. However, it has been admitted that the overall performance concerning precision of YOLO is lower as compared to faster R-CNN, mainly due to complicated small object detection and a false positive tendency in result generation [43]. The YOLO algorithm has been improved in an iterative manner, and further versions have shown better performance compared to the former ones [44]. The YOLOv2 (YOLO9000) introduced the anchor boxes to handle the size of the objects better [44]. On top of this model, YOLOv3 put in a multi-scale recognition mechanism on the Darknet-53 [44]. Later, with YOLOv4 and YOLOv5, different advanced techniques like CSPNet, PANet, and Mish activation were used to make the model optimal for more speed and accuracy [44].

YOLO is a unique method of object identification, with the objects' classifications and localizations well merged. It is very efficient since it has the unique ability to assess an entire image in just a single pass, making it very fast [42]. In the field of neural network structures, different structures have been developed to handle different tasks. The Convolutional Neural Network (CNN), which is specialised in image processing, is the best at detecting a pattern in an image through its architecture; hence, it can perform well on image recognition, text detection, colour detection, describing an image, and object detection [45]. In the case of sequential data, such as time series or language patterns, two big classes of structures are represented by the Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) models. The LSTM is a subclass of RNN designed to solve the "vanishing gradient problem" present in the original RNN architecture [38]. This enhancement in the LSTM architecture makes it better for learning long-range dependencies in the sequence data, making it particularly well-suited for temporal and sequence information-dependent tasks.

**Table 2: Various AI Technologies.**

| Name of AI technology | Image recognition | Text detection | Colour detection | Object recognition | Describes image | Object detection | NLP |
|---|---|---|---|---|---|---|---|
| CNN | * | * | * | | * | * | |
| ResNet-50 | | | | * | | | |
| YOLO | | | | * | | * | |
| RNN | | | | | | | * |
| LSTM | | | | | | | * |
| MobileNet-SSD | | | | * | | * | |

The Table 2 above provides a summary of different AI technologies and their capabilities as seen below. Convolutional neural networks (CNNs) evolved as the primary architecture in deep learning in relation to processing and understanding visual input. Their built-in ability to see and understand spatial arrangements makes CNNs versatile in an array of tasks, including image and colour recognition and object detection. The ResNet-50 architecture really marks a big leap for CNNs in depth and introduces residual connections. These features are intended to solve the problem of vanishing gradients, making them more effective in the tasks that complex object identification involves. On the other hand, the YOLO framework has been very popular as it is highly efficient when used in real-time object detection. It followed a special technique of grid-based classification applicable to the fast detection of several items from visual frames. In sequential data, recurrent neural networks and their more advanced versions, long short-term memory networks, are somewhat different from each other. This makes them inherently recursive and well-suited to tasks with sequential understanding, such as text identification and

complex natural language processing efforts. The combination of MobileNet and SSD is very efficient in the case of constrained computational resources. This concoction has enabled it to strike a trade-off between efficiency and accuracy, making it very apt for on-device visual processing scenarios. The effectiveness of the system, along with the ability of the user, must be accessible and comprehensible in a way that ensures the user is efficient in the long term in the ensuing interactions.

## 2.4 ISO Standards of Ergonomics Design

ISO 9241-110:2006, Ergonomics of Human-System Interaction, Part 110: Interaction principles supply demands and pragmatic principles for decorating discussions amid public and in-person systems. The standards apply to the constituents of an exchange where there are interfaces between people, equipment, and software, including both hardware (such as screens, buttons, and keyboards) and software (such as dialogue boxes and menus).

### 2.4.1 ISO 9241-110 lists seven general principles for effective user-interface design [46]:

**Suitability for the task**: The interface should prioritize assisting users in completing relevant tasks, promoting efficiency and effectiveness in their interactions.

**Self-descriptiveness:** The design should ensure that each step of user interaction is inherently understandable, achieved through intuitive design or the timely supply of explicit information.

**Controllability:** Users should have the capacity to initiate, control, and choose the pace of interaction until the desired goal is reached.

**Conformity with user expectations:** Consistent conduct and adherence to acknowledged norms and standards are crucial for predictability, aligning the system with user expectations.

**Error tolerance:** The system should have a robust architecture that allows easy correction when users make mistakes, ensuring that desired outcomes can still be achieved.

**Suitability for individualization:** The system should be adaptable to accommodate a diverse range of users, allowing for adjustments that align with individual preferences or requirements, leading to a customized user experience.

**Suitability for learning:** The system should ensure accessibility and comprehensibility regardless of the user's expertise, fostering sustained user efficiency during recurring interactions.

These principles are designed for systems that are more usable and provide users with an overall better experience. Though not focused on visual impairment, the design principles are claimed to be applicable for any human-computer interaction. Systems conforming to such principles are expected to be more accessible and user-friendly to a wide range of users, including those with visual impairments, with better access and user interaction with the system. Following is a primary investigation of secondary research, and questions are designed by considering the ergonomics of human-computer interaction as given by the guidelines of ISO.

# 3 Methodology And Study Design

In a study on the effectiveness of current image recognition systems towards enabling digital image understanding by a person with visual impairments, firsthand information from the intended user group was very important. This included the perceptions and experiences of visually impaired people. We solicited some demographic information about the individuals working with image recognition tools, such as their level of education, professional background, and experience with image recognition tools. The findings have provided insights that helped to understand, on a higher level, the way the visually impaired interact with image recognition tools. Such knowledge is really important in devising strategies for improved accessibility and effectiveness [47]. This survey data helps identify the strengths and weaknesses of tools for image recognition, the potential ethical considerations, and strategies devised for increasing their accessibility and overall effectiveness.

The following survey questions were designed to capture the current state of image recognition tools.

1) Do you use any technology to understand images?
2) If “Yes”, what is it? If “No”, how do you understand digital images?
3) Do you trust this tool?
4) Does this tool accurately describe the image you think?
5) Does this tool sufficiently describe the image you think?
6) What do you like about these tools?
7) What don’t you like about these tools?
8) Would you recommend your tool to another blind person?
9) How do you like a digital image to be explained by an Image Recognition tool?
10) Would voice feedback of an image be sufficient?
11) If not, what other feedback mechanism would you like?
12) What features would you like to see in an Image Recognition tool for digital image recognition?
13) Can you rate the importance of the following features of an Image Recognition tool? Voice results, Object detection, Colours detection, Light detection, Voice Commands, Face analysis, Reads handwriting, Currency recognition, Describes environment.

The design of the survey questions focused on descriptive analysis, which thereby allows for the summarization of the received data towards the identification of basic insights and patterns. 'Yes/No' questions, such as questions 1, 3, 4, 5, 8, and 10, elicited the count or percent that emanated from the answers 'Yes' and 'No'. In Question 2, where the respondent was to detail the type of technology in use, a categorization of the responses was carried out, thereby offering a count or percentage of each of the categories as a category analysis. Questions 6, 7, 9, 11, and 12 were designed to be free responses and expected to have lengthy details. These responses will be analysed through content analysis, a qualitative method that categorises text data into groups. This allows for the content to be looked at, thematic groups generated, and the number of responses in that category.

In relation to Question 13, the importance of the available features is calculated using means and standard deviations in order to assess how significant the feature is to the respondent. In short, it is an attempt to give a detailed explanation of how the visually impaired people use the image recognition tools, with particular respect to their perception of these tools, the level of satisfaction obtained, and the expectations of such technology. Ethical approval in any research concerning an individual with a visual impairment is important so that the study is ethically and responsibly conducted, and their rights, safety, and well-being are protected within the confines of the research under consideration. The current study also got ethical clearance through the Ethics Review Board of the institution. The board shall look into the study design, research protocols, informed consents, and participant information sheet, all in light of conformity with the norms and guidelines of ethical approval. Special considerations should be taken into account while conducting research on such people, like how they navigate novel environments or gain access to printed materials, and thus, ethical approval ensures that a visually impaired person will be adequately informed about the study and able to give consent to it. It also ensures that a person can ask questions and obtain information in a manner that is responsive to his needs in a given environment. “Ethical approval assures that the potential benefits of the research must outweigh the possible risks or harm to the participants; ergo, it maintains this balance throughout the course of the study.".

## 3.1 Procedure

The research process with the visually impaired required various key steps. Firstly, the target population, including those with complete blindness or low vision, was defined precisely. Secondly, outreach was extended to a number of organisations and charities focused on the visually impaired community to enlist their assistance in connecting with potential participants for the survey and in sharing information about the survey. Thirdly, recruitment methods are committed to the accessibility of communication through the use of plain language and the delivery of materials in accessible formats for screen readers or other assistive technologies being used. Fourthly, clarity in communication was a priority, giving assurance that participants understood the purpose of

the survey. In addition, consent forms and information for participants were made available in an accessible format.

Finally, meetings were scheduled with participants, and questions for the survey, which were previously prepared, were circulated in accessible formats in advance of the interview to assist with their preparation. Together, these measures ensure that, in essence, ethical approval is taken as a given component and that research is competently and ethically conducted with an eye towards the good of participants and the protection of their rights and well-being. It also assists in coming up with valuable insights regarding the experiences of people with visual impairments. Table 3 summarises the demographic information of the participants, including age, gender, education level, computer literacy, age the participant was when they became blind, and the cause of the participants' blindness. The responses from the participants included in this study were ten, who were part of an image recognition experiment and were each given a unique identification code for recording purposes. Of the ten participants, six were males and four were females. Notably, all participants had post-secondary education at the tertiary level. A notable observation is that only one participant reported having limited proficiency in computer literacy.

Concerning visual capacities, four members of the sample lost their vision completely within several months after birth; the lapse of time from birth to the loss of vision was less than one year. The small residual vision was kept by half of the sample—five people—whereas the other half did not have residual vision. Most of them attribute the cause of their visual impairment to a number of factors, with genetic causes being the sole factor in most of the cases. The rest of these demographic details give an overview of the participants' profiles in this research study.

***Table 3:*** *Participant's Demographics.*

| Participant code | Age (yrs) | Sex | Education | Computer literacy | Age of blindness (yrs) | Residual vision | Reason for blindness |
|---|---|---|---|---|---|---|---|
| P 1 | 41 | Male | Tertiary | Substantial | 9 | Yes | Genetic issues |
| P 2 | 49 | Female | Tertiary | Average | 11 | Yes | Stargardt |
| P 3 | 71 | Male | Tertiary | Substantial | < 1 | No | Infection |
| P 4 | 69 | Female | Tertiary | Average | 15 | No | Retina issues |
| P 5 | 74 | Male | Tertiary | Average | 2 | No | Smallpox |
| P 6 | 66 | Female | Tertiary | Average | 18 | Yes | Iris issues |
| P 7 | 32 | Male | Tertiary | Average | 16 | No | Retina issues |
| P 8 | 44 | Male | Tertiary | Average | < 1 | Yes | Retina issues |
| P 9 | 33 | Male | Tertiary | Substantial | < 1 | No | Premature development |
| P 10 | 60 | Female | Tertiary | Minimal | < 1 | Yes | Optic nerve issues |

# 4 Finding (Qualitative and Quantitative)

Almost all of the participants were happy with the voice feedback they received but wished for item detection, face analysis, and environment description. Opinions about colour and light detection vary. Although a few issues were identified, most had a positive disposition to recommend the device to other peers with visual impairments.

This section explains data collected from the interviews and surveys.

- **Technology usage - Do you use any technology to understand images?**

Every participant noted using various technological tools that support the reading of visual material. The tools included the auto-recognition features of Twitter and iPhone, Seeing AI, Be My Eyes, Sullivan+ on iPhone, TapTapSee, and JAWS. This variety only indicates how extensive the range of tools on the market is, intended to

support the visually impaired population in attaining and interpreting visual information. Figure 1 below shows the summary of the usage of some of the assistive technologies.

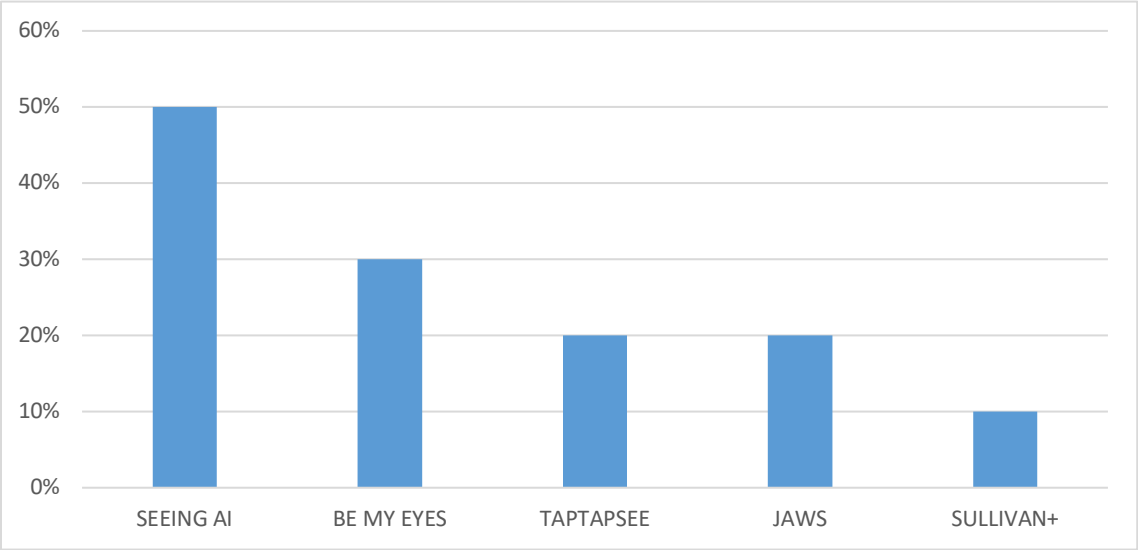


**Figure 1: Usage of five assistive technologies.**

- **Trust and Accuracy - Do you trust this tool? Does this tool accurately and sufficiently describe the image you think?**

A range of trust can be found among participants based on the image recognition techniques employed, with a majority seeming to be taking a positive sentiment in answering their trust as either "yes" or "partially." This means that there is still room for improvement, but still, a good number of visually impaired people are confident with the choice of assistive technology made. Most of them further expressed their confidence in the image recognition capabilities being right and sufficient to describe the images, although some rare exceptions were stated. This means that the overall result from such technologies works out to what the user expects but might not be perfect in all cases. Figure 2 below summarises the Trust and Accuracy data collected.

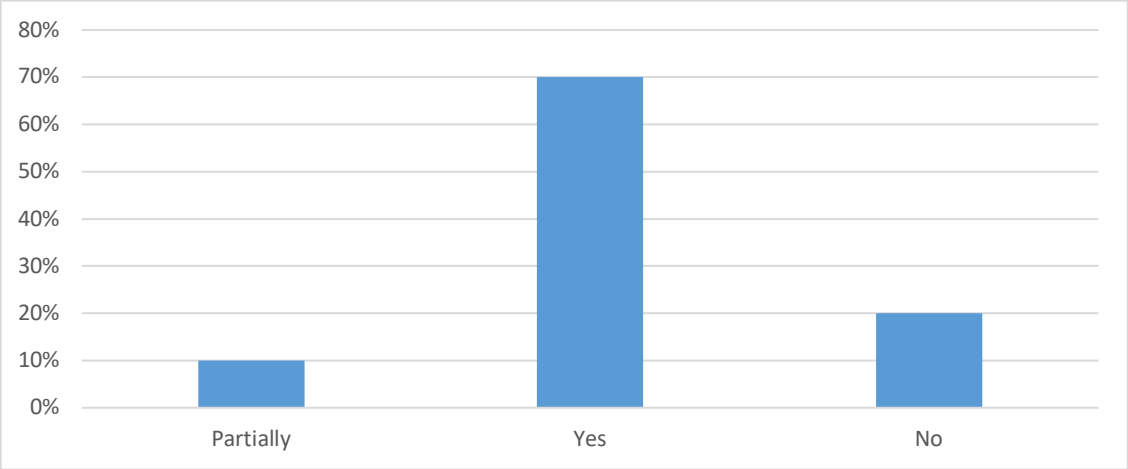


**Figure 2: Trust and accuracy of measured by blind users**

- **Likes of assistive technology - What do you like about these tools?**

The participants appreciated the ease of access to the tools provided by their smartphones and the fast delivery of descriptions. In this way, such feedback has the potential to be very useful towards the development of designs and implementations for image recognition technologies. Figure 3 below summarises four features which blind and cisually impaired people likes most.

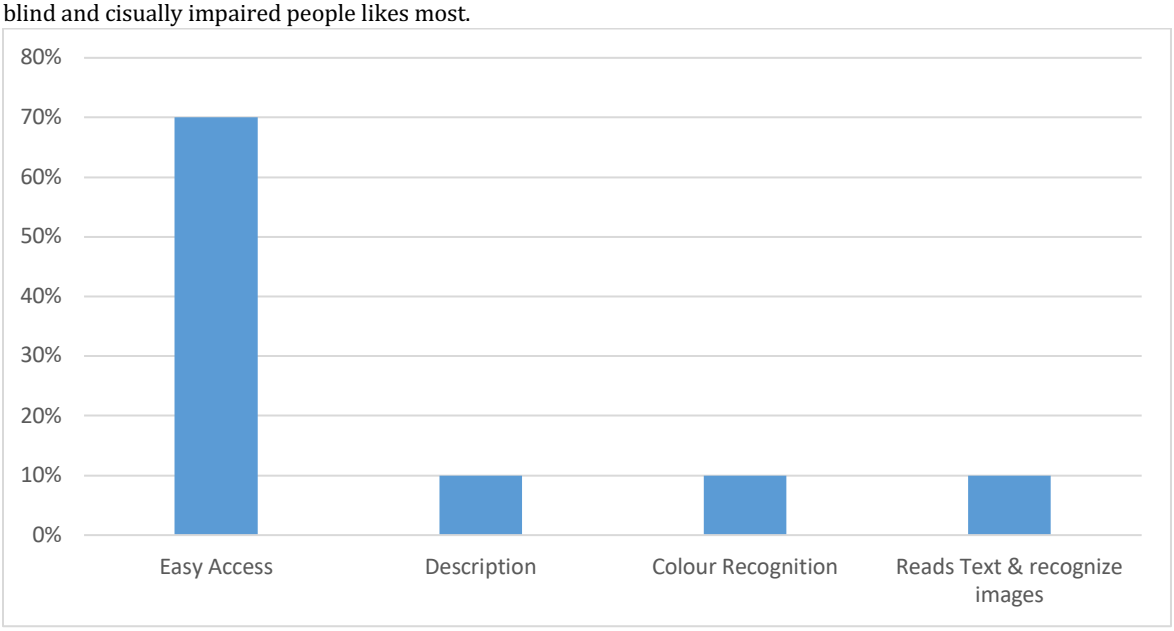


**Figure 3: The percentage of blind users that likes four key features of assistive technology.**

- **Dislikes of assistive technology - What don't you like about these tools?**

They expressed dissatisfaction in different aspects of the app: integration is not smooth; one has to be sure of the camera's position without seeing the camera feedback; there are security concerns; the customisation of

options is not sufficient; and the image descriptions provided did not have enough detail. Figure 4 below summarises the features that are disliked by the blind and visually impaired people.

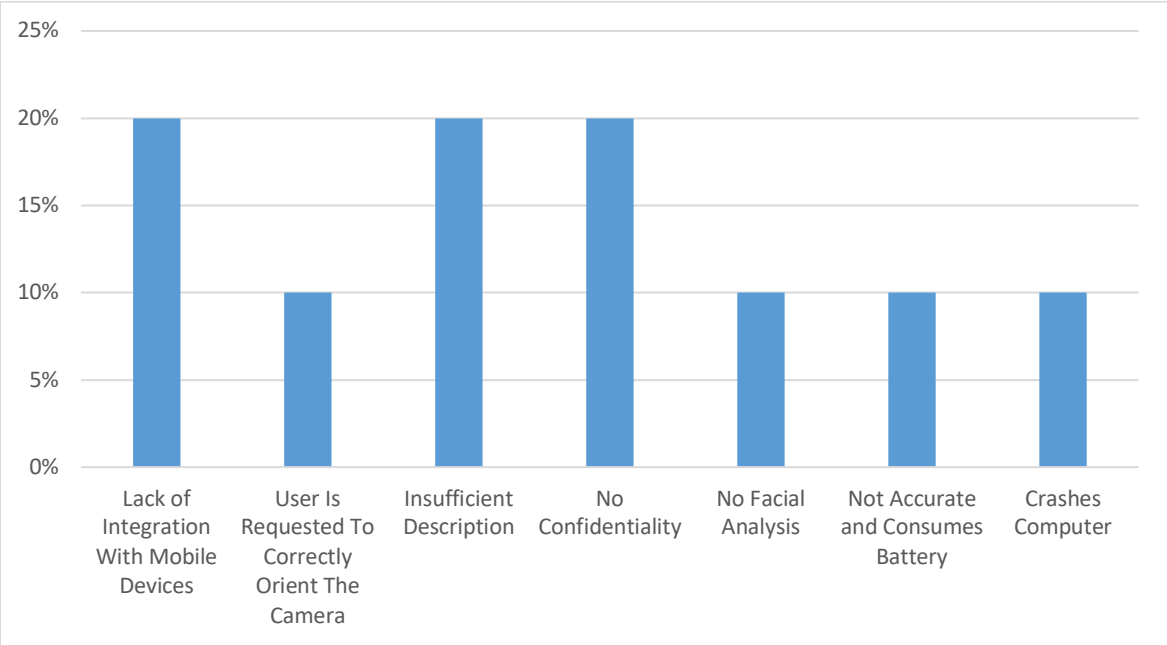


**Figure 4: Displays the percentage of visually impaired users who express dissatisfaction with seven key features of assistive technology highlighted in this study.**

- **Desired features - How do you like a digital image to be explained by an Image Recognition tool?**

They felt the need for some more added features with the image recognition systems, such as improved descriptive capabilities, increased accuracy, and auditory elements incorporated in the system to improve the overall user experience. Figure 5 below summarises the feature desired by blind and visually impaired people.

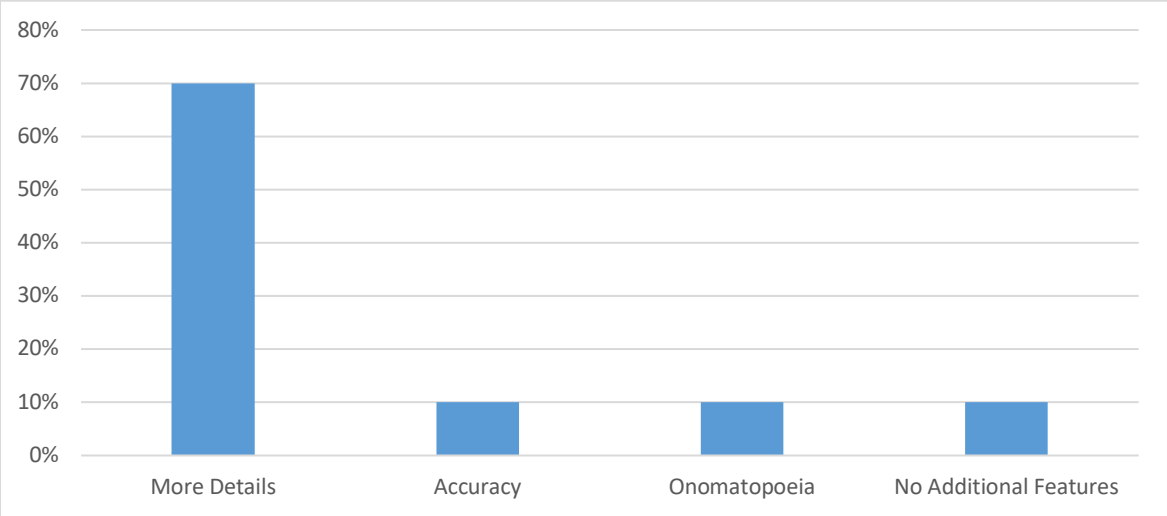


**Figure 5: Desired features of assistive technology.**

- **Tool recommendations - Would you recommend your tool to another blind person?**

All participants claimed they would recommend their selected image recognition tool to other people with visual impairments.

- **Description preferences - How do you like a digital image to be explained by an Image Recognition tool?**

Some participants called for more detailed visual descriptions, while others wanted less explicit descriptions, especially for outdoor settings. It shows that customisable options would cater to individual preferences.

- **Voice feedback - Would voice feedback of an image be sufficient?**

Every single one agreed that the voice feedback provided by the systems of image recognition was satisfactory without any exception, which proves the efficiency of such a strategy in the successful delivery of information.

- **Feature importance ratings - What features would you like to see in an Image Recognition tool for digital image recognition?**

  In fact, they tested the importance of different attributes regarding the techniques used in image recognition: voice outcomes, object detection, face analysis, and environment description. The results came out quite high, and it turned out that the functions are really valued by consumers. Figure 6 below shows the summary of the outcomes highlighted earlier.

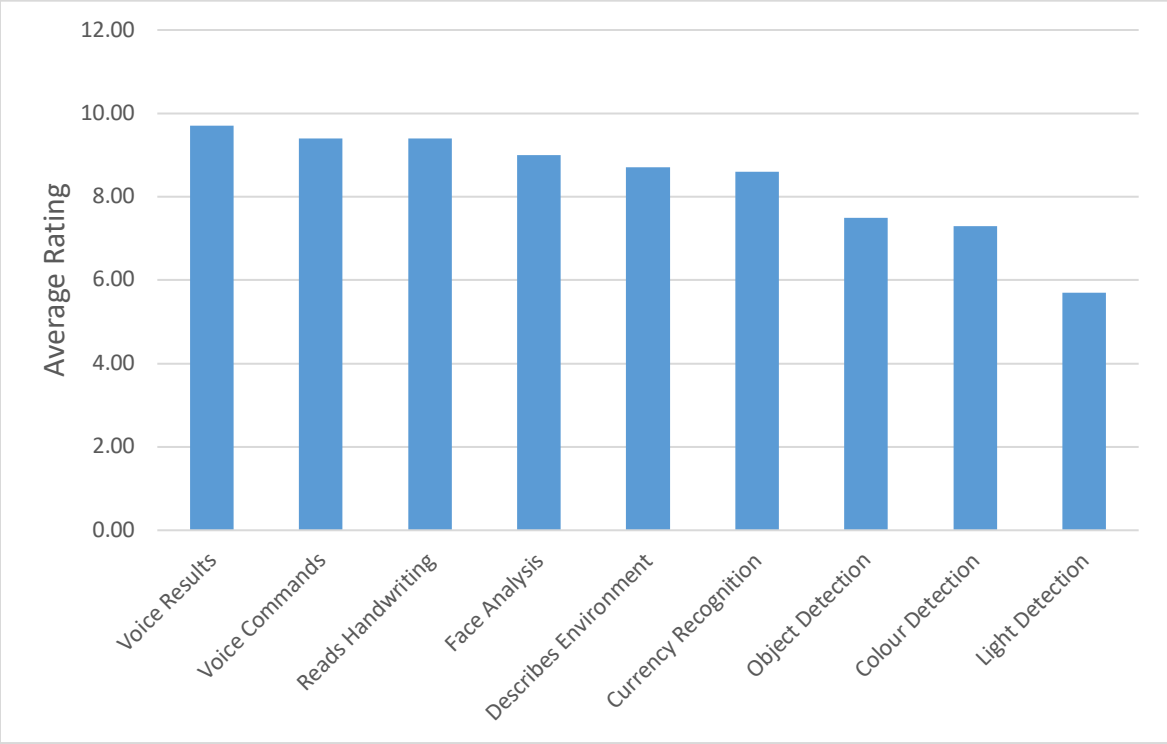


**Figure 6: The percentage rating by blind and visually impaired people.**

In summary, image recognition technology plays a crucial role in the lives of visually impaired individuals, enhancing their access to information and promoting autonomy. However, there is potential for improvement in terms of precision, integration, and functionalities. Future studies should focus on developing more sophisticated image recognition technologies to address these objectives. There is room for improvement in some IR features and a need for the integration of others. One of the limitations of this study is the number of participants. For a complete comprehensive study, future research should consider candidates from all unique forms of disability and provide sufficient time to study all the systems mentioned in Table 1, which can be costly and take longer time. Blind users may easily spend more time getting familiar with systems and can report different satisfaction levels. Data from separate groups should be analyzed separately, and a more comprehensive description of participants should be recorded for experiment replication purposes. An alternative approach that is used in this study is to access somewhat random diverse people in blind community and use standard statistical methods to state results without generalizing user opinions.

# 5 DISCUSSION

An essential inquiry lies in assessing the accuracy of the underlying algorithm. Should it prove inaccurate, should we pivot towards employing new data and retraining the algorithm? Concurrently, attention must be paid to the Human-Computer Interaction (HCI) layer overlaying Deep Learning algorithms in app development. The interaction dynamics differ significantly for blind individuals compared to sighted users, necessitating a strategic alignment of mobile application objectives with the activity sequence of blind users. This alignment, guided by

activity theory [48], seeks to enhance the design of nonvisual computing systems, ultimately reducing barriers to entry for the blind.

Feedback from blind users underscores dissatisfaction with the precision, integration, and functionalities of existing systems. Future endeavours should prioritize the advancement of picture recognition technologies to address these concerns comprehensively. Furthermore, a broader and more diverse exploration of user experiences is imperative for a nuanced understanding of the subject while this study investigates the user interaction layer of IR tools.

While existing image recognition tools for individuals with visual impairments partially reflect ISO features outlined in section 2.4, there's room for enhancement and integration of additional functionalities. Categorized study findings offer insights into IR developers' utilization, features, suitability, descriptiveness, tolerance, individualization, and suitability for learning.

Among the study participants, certain tools enjoy widespread usage: 'Seen AI' (50%), 'Be My Eyes' (30%), 'TapTapSee' (20%), 'JAWs' (20%), and 'Sullivan' (10%).

Key image recognition features deemed significant include voice feedback (100%), object detection (80%), colour detection (80%), light detection (60%), voice commands (100%), face analysis (80%), handwriting recognition (90%), currency recognition (90%), and environment description (85%).

Regarding suitability and descriptiveness, 70% of participants expressed confidence in the accuracy and adequacy of their image recognition capabilities, with minor exceptions. Notable features highlighted include easy accessibility (70%), level of description (10%), color prediction (10%), and text and image recognition (10%).

In terms of controllability, conformity, and error tolerance, certain desired features lacked essential attributes, as reported by participants: lack of mobile device integration (20%), necessity for precise camera orientation (10%), insufficient description (20%), lack of system confidence (20%), absence of facial analysis (10%), inaccuracy (10%), high battery consumption (10%), and system crashes (10%).

In terms of qualitative feedback, we learnt that participants articulated keen interest in augmenting their image recognition systems with added functionalities, notably emphasizing the need for customizable descriptive features, heightened precision, and the incorporation of auditory elements to enrich user experiences. This expansion is anticipated to bolster individualization and aptness for learning. Additionally, participants advocated for the integration of image recognition systems with supplementary modalities to amplify flexibility, accessibility, and usability. Such integration is poised to cater to diverse cognitive processing styles and requirements. Embracing multimodal approaches, image recognition systems can elevate user experiences, enhance task performance, and adeptly accommodate individual preferences and capabilities. It's essential, however, to incorporate these features as optional outputs aligned with user preferences, thus circumventing information overload.

While the algorithms and detailed development processes outlined in Table 1 remain undisclosed to the public, the design and development of these systems typically encompass several crucial features:

- Text-to-speech functionality.
- Screen reading capability.
- Accessibility API facilitating access to UI elements within the content.
- Navigation and focus management enabling seamless navigation across documents, web pages, and applications through straightforward navigation, clear instructions, and intuitive controls.
- Customization options and user preferences, including speech output, navigation preferences, and other settings, ensuring usability across various impairment levels (supporting screen readers, voice commands, and adjustable font sizes).
- Compatibility and integration with diverse drivers and software environments.

These technologies come with associated drawbacks, primarily concerning accuracy, speed, and accessibility:

**Accuracy:** Despite advancements, these technologies may struggle with accuracy, especially in complex or cluttered environments. They might misclassify objects or scenes, leading to interpretation errors. Machine

learning algorithms, such as CNNs, heavily rely on the quality and diversity of training data. Biased or incomplete training data can result in inaccuracies and limitations in recognizing and interpreting visual information. While proficient at recognizing patterns and extracting features, these algorithms may face challenges in understanding contextual information, potentially leading to misinterpretations. Mobile applications for visual assistance must operate in dynamic, real-time environments. CNNs and object detection algorithms may encounter difficulties with rapidly changing scenes or objects, causing delays or inaccuracies in providing assistance.

**Speed**: Mobile devices often have limited processing power compared to desktop computers or servers, impacting the speed and responsiveness of applications relying on resource-intensive algorithms like CNNs for image processing or NLP for text recognition.

**Accessibility:** While these technologies aim to aid people with visual impairments, they may encounter accessibility barriers for users with varying levels of impairment or those utilizing assistive technologies other than screen readers. Mobile applications relying on CNNs and NLP often necessitate access to the device's camera and microphone, raising privacy concerns regarding data collection and storage. Ensuring user privacy while utilizing these technologies is crucial.

BVI system designers and developers face the imperative of overcoming limitations through advancements in algorithmic techniques, data collection, and user-centred design to forge more potent and inclusive solutions. The study's findings furnish pivotal insights, underscoring that while image recognition tools furnish some assistance to individuals in comprehending digital images, significant scope for enhancement persists. This underscores the necessity for further research and innovative development endeavours to bridge extant gaps in usability and efficacy.

While tools like Seeing AI and JAWS distinguish themselves in suitability for the task, self-descriptiveness, and controllability, others such as AI smart glass and Audible Vision, though task-relevant, may necessitate enhancements in controllability and error tolerance. Tools like Amazon Rekognition and Roboflow, primarily aimed at developers, could be refined to better cater to the visually impaired by aligning with ISO principles. Meanwhile, tools like AIpoly, Seeing AI, and LookTel, crafted to aid the visually impaired, exhibit promising functionalities, ranging from real-time object identification to currency recognition, evincing potential for augmenting digital accessibility.

To holistically assess the efficacy of these tools, it is imperative to evaluate them in congruence with ISO standards, incorporating crucial elements such as suitability for the task, self-descriptiveness, controllability, and error tolerance, alongside other pertinent criteria. The crux of this article lies in dissecting ISO standards and elucidating their relevance to blind users, thereby engendering more precise design implications.

# 6 CONCLUSION

Our study reveals the necessity for the next generation of tools to be not only technologically advanced but also harmonized with user-centric standards like those delineated by ISO. Current publications rarely provide a comprehensive evaluation of common design mistakes or user access barriers in image recognition (IR) tools for blind individuals. Instead, they tend to introduce new systems, highlighting their strengths, weaknesses, and offering some analysis of the current state of the art. This paper, however, adopts a more holistic approach by examining a diverse range of IR tools from various sources. It critiques these tools through the lens of front-end user feedback and backend design techniques, delivering a detailed analysis while staying focused on the specific needs of blind users. It's imperative to integrate user feedback and conduct extensive usability testing to realize this objective. Drawing on insights gleaned from our survey and research, an optimal image recognition tool tailored for visually impaired users should embody the following key characteristics: (1) Inherent Suitability: The tool must be naturally designed for visually impaired individuals, ensuring that its fundamental functionalities align with their specific level of disability and requirements. It should be readily usable by blind individuals without significant customization. (2) Intuitive User Interface: Given the unique challenges faced by these users, the tool should feature a self-descriptive, intuitive user interface to enhance accessibility. This interface should

consider the familiar sequence of digital activities of the target users. (3) Customization: Recognizing the variability in visual impairments and user preferences, the tool should allow for individual customization, catering to diverse needs. (4 ) Error Minimization and Correction: The IR tool should not only be tolerant of errors but should also provide mechanisms for users to easily rectify mistakes, thereby enhancing overall usability. (5) Continuous Learning: Leveraging artificial intelligence, the tool should actively learn from user interactions, continuously refining its responses and predictions over time to provide an improved user experience.

While current image recognition tools offer promise in revolutionizing digital accessibility for the visually impaired, there's a crucial need to prioritize a user-centric approach. Future advancements should seamlessly integrate technological capabilities with user feedback to ensure these tools authentically meet the needs of their primary audience. This study not only contributes to the existing knowledge base but also provides a practical roadmap for advancing digital inclusivity.

Looking ahead, our research will focus on developing a suitable prototype guided by the outlined principles. This involves refining computer vision algorithms for various types of image detection, improving the accessibility features of image recognition tools for the blind, and conducting an extensive, long-term multi-user study of an image recognition prototype. This endeavour will involve utilizing appropriate APIs to identify areas for improvement and further enhancing the overall effectiveness and usability of image recognition tools for the visually impaired.

## ACKNOWLEDGMENTS

We extend our sincere thanks to the Assistive Technology Group and Communication Technology Research Centres at London Metropolitan University for their invaluable guidance and support. Special gratitude is also extended to the visually impaired participants from Baluji Music Foundation, Beyond Sight Loss, and RNIB who generously shared their experiences.

## REFERENCES

<bib id="bib1"><number>[1]</number>Adam Zewe, Making data visualization more accessible for blind and low-vision individuals, 2022, https://news.mit.edu/2022/data-visualization-accessible-blind-0602.</bib>
<bib id="bib2"><number>[2]</number>Alexiou Gus, Envision Smart Glasses – A Game-Changer In Helping Blind People Master Their Environment, 28 01 2021, Forbes, https://www.forbes.com/sites/gusalexiou/2021/01/28/envision-ai-glasses--a-game-changer-in-helping-blind-people-master-their-environment/.</bib>
<bib id="bib3"><number>[3]</number>Al-Qunaieer, F.,, 2014. Automated Resolution Selection for Image Segmentation. s.l.:s.n.</bib>
<bib id="bib4"><number>[4]</number>Andy Nguye, Liamputtong, P. (eds) Handbook of Social Inclusion In: Digital Inclusion, 2021, Springer, https://link.springer.com/referenceworkentry/10.1007/978-3-030-48277-0_14-1#citeas, https://doi.org/10.1007/978-3-030-48277-0_14-1.</bib>
<bib id="bib5"><number>[5]</number>Antol, S., Agrawal, A., Lu, J., Mitchell, M., Zitnick, C. L., Batra, D., & Parikh, D. (2015). VQA: Visual Question Answering. arXiv preprint arXiv:1505.00468, https://arxiv.org/abs/1505.00468.</bib>
<bib id="bib6"><number>[6]</number>AppAdvice, Aipoly Vision: Sight for Blind & Visually Impaired, 2017, 13 12 2022, https://appadvice.com/app/aipoly-vision-sight-for-blind-visually-impaired/1069166437.</bib>
<bib id="bib7"><number>[7]</number>Babbie, E.R. (2016) The Practice of Social Research. 14th Edition, Cengage Learning, Belmont, https://www.scirp.org/reference/ReferencesPapers.aspx?ReferenceID=2439585.</bib>
<bib id="bib8"><number>[8]</number>Baldwin, Mark S. and Mankoff, Jennifer and Nardi, Bonnie and Hayes, Gillian, 2020. An Activity Centered Approach to Nonvisual Computer Interaction. New York (NY): ACM, https://doi.org/10.1145/3374211.</bib>
<bib id="bib9"><number>[9]</number>Banerjee, S., Deb, K., Das, A. and Bag, R., 2021. In Handbook of Research on Modern Educational Technologies, Applications, and Management. In: A Survey on the Use of Adaptive Learning Techniques Towards Learning Personalization. s.l.:IGI Global, pp. 790-808.</bib>
<bib id="bib10"><number>[10]</number>Bhanuka Gamage, Thanh-Toan Do, Nicholas Seow Chiang Price, Arthur Lowery, Kim Marriott, 2023. What do Blind and Low-Vision People Really Want from Assistive Smart Devices? Comparison of the Literature with a Focus Study. New York, NY, ACM, https://doi.org/10.1145/3597638 .</bib>
<bib id="bib11"><number>[11]</number>Bill Holton, A Review of the TapTapSee, CamFind, and Talking Goggles Object Identification Apps for the iPhone, 2013, American Foundation for the Blind, https://www.afb.org/aw/14/7/15675.</bib>
<bib id="bib12"><number>[12]</number>Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M., 2020. YOLOv4: Optimal speed and accuracy of object detection. s.l.:arXiv preprint, https://arxiv.org/pdf/2004.10934.pdf.</bib>
<bib id="bib14"><number>[14]</number>Cabrera, Á.A., Perer, A. and Hong, J.I., 2023. Improving human-AI collaboration with descriptions of AI behavior. Proceedings of the ACM on Human-Computer Interaction. s.l., Association for Computing Machinery, pp. 1-21.</bib>
<bib id="bib15"><number>[15]</number>Chandrika Jayant, Hanjie Ji, Samuel White, Jeffrey P. Bigham, Supporting Blind Photography, 2011, ACM, Dundee, Scotland, UK., https://cs.rochester.edu/hci/pubs/pdfs/supporting-blind-photography.pdf.</bib>

<bib id="bib17"><number>[17]</number>Gamage, Bhanuka, Thanh-Toan Do, Nicholas Seow Chiang Price, Arthur Lowery, and Kim Marriott, 2023. What do Blind and Low-Vision People Really Want from Assistive Smart Devices? Comparison of the Literature with a Focus Study.. Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 1-21.</bib>
<bib id="bib18"><number>[18]</number>Girshick R., Donahue J., Darrell, T. & Malik J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. s.l., IEEE, pp. 580-587, https://arxiv.org/abs/1311.2524.</bib>
<bib id="bib19"><number>[19]</number>Hayashi, T., Cimr, D., Fujita, H. and Cimler, R., 2023. Image entropy equalization: A novel preprocessing technique for image recognition tasks. s.l.:Information Sciences.</bib>
<bib id="bib20"><number>[20]</number>Hersh, M.A, 2018. Pissaloux, E., Velazquez, R. (eds) Mobility of Visually Impaired People.. In: Mobility Technologies for Blind, Partially Sighted and Deafblind People: Design Issues.. s.l.:Springer, Cham., p. 377–409.</bib>
<bib id="bib21"><number>[21]</number>Hinton, G. E., Sutskever, I., & Krizhevsky, A., 2012. ImageNet classification with deep convolutional neural networks, arXiv:1207.0580v1.</bib>
<bib id="bib22"><number>[22]</number>Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.. s.l.:arXiv preprint, http://export.arxiv.org/abs/1704.04861.</bib>
<bib id="bib23"><number>[23]</number>IEEE, 2023. IEEE Standard for Robustness Testing and Evaluation of Artificial Intelligence (AI)-based Image Recognition Service. IEEE Std 3129-2023, 2 June.pp. 1-34</bib>
<bib id="bib24"><number>[24]</number>ISO, ISO 9241-110:2006 Ergonomics of human-system interaction — Part 110: Dialogue principles, 2006, 19 September 2023, https://www.iso.org/standard/38009.html.</bib>
<bib id="bib25"><number>[25]</number>ISO, ISO 9999:2016 - Assistive products for persons with disability — Classification and terminology, 2018, ISO, https://www.iso.org/standard/50982.html.</bib>
<bib id="bib26"><number>[26]</number>ISO, ISO/IEC 20071-21:2015 - Information technology — User interface component accessibility — Part 21: Guidance on audio descriptions, 2019, International Organization for Standardization, https://www.iso.org/standard/70485.html.</bib>
<bib id="bib27"><number>[27]</number>J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.</bib>
<bib id="bib28"><number>[28]</number>Jack Lenstrohm, TapTapSee App Review, 12 July 2019, 13 12 2022, blindresources, https://blindresources.org/2019/07/12/taptapsee-app-review/.</bib>
<bib id="bib29"><number>[29]</number>K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.</bib>
<bib id="bib31"><number>[31]</number>Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C., 2016. European conference on computer vision. In: SSD: Single shot multi-box detector. s.l.:Springer, Cham., pp. 21-37, https://arxiv.org/abs/1512.02325.</bib>
<bib id="bib32"><number>[32]</number>Liu, Y.H., 2018. Feature extraction and image recognition with convolutional neural networks. Journal of Physics: Conference Series, September, Volume 1087, p. 062032.</bib>
<bib id="bib33"><number>[33]</number>LookTel, LookTel Recognizer, 2009, 11 12 2022, https://www.iaccessibility.com/apps/low-vision/index.cgi/product?ID=44.</bib>
<bib id="bib34"><number>[34]</number>LookTel, What is LookTel?, 2009, 16 9 2023, http://www.looktel.com/.</bib>
<bib id="bib35"><number>[35]</number>Luo, M. and Zhang, K., 2014. Engineering Applications of Artificial Intelligence. In: A hybrid approach combining extreme learning machine and sparse representation for image classification. s.l.:ScienceDirect, pp. 228-235.</bib>
<bib id="bib36"><number>[36]</number>Marcus Klasson and Cheng Zhang, Using Variational Multi-view Learning for Classification of Grocery Items, 2020, CellPress, 13 12 2022, https://www.cell.com/patterns/pdfExtended/S2666-3899(20)30191-4.</bib>
<bib id="bib38"><number>[38]</number>Microsoft, How Seeing AI works, 18 February 2023, https://www.microsoft.com/en-us/seeing-ai/how-it-works.</bib>
<bib id="bib39"><number>[39]</number>OpenAI, 2023. GPT-4V(ision) System Card. s.l.:s.n.</bib>
<bib id="bib40"><number>[40]</number>Raz Ben, Aipoly Reviews - Pros & Cons 2022 - Product Hunt, 2017, 13 12 2022, https://www.producthunt.com/products/aipoly/reviews.</bib>
<bib id="bib41"><number>[41]</number>Redmon, J., & Farhadi, A., 2018. YOLOv3: An incremental improvement. s.l.:arXiv preprint, https://arxiv.org/pdf/1804.02767.pdf.</bib>
<bib id="bib42"><number>[42]</number>Robert Sanders, Digital inclusion, 2020, Iriss, https://www.iriss.org.uk/sites/default/files/2020-04/iriss_esss_outline_digital_inclusion_09042020_0.pdf.</bib>
<bib id="bib43"><number>[43]</number>S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.</bib>
<bib id="bib44"><number>[44]</number>S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.</bib>
<bib id="bib45"><number>[45]</number>Saeed, S., Rohde, M., & Wulf, V. (2013). Designing an Assistive Learning Aid for Writing Acquisition: A Challenge for Heterogeneous Groups. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2033-2042), https://www.scirp.org/journal/paperinformation.aspx?paperid=123700.</bib>
<bib id="bib46"><number>[46]</number>Sanni Siltanen, Theory and applications of marker-based augmented reality, 2012, Vuorimiehentie, Finland, JULKAISIJA – UTGIVARE – PUBLISHER, *https://www.vttresearch.com/sites/default/files/pdf/science/2012/S3.pdf*.</bib>
<bib id="bib49"><number>[49]</number>Sutskever, I., Vinyals, O., & Le, Q. V., 2014. Sequence to sequence learning with neural networks, arXiv:1409.3215.</bib>
<bib id="bib50"><number>[50]</number>Teresa Correa, Isabel Pavez, Digital Inclusion in Rural Areas: A Qualitative Exploration of Challenges Faced by People From Isolated Communities In: Journal of Computer-Mediated Communication, 2016, 247–263, 21, 3, *https://academic.oup.com/jcmc/article/21/3/247/4065369*.</bib>
<bib id="bib51"><number>[51]</number>Tushar P Ghate, Sukruti Y Khairnar, Santosh A Bangar, J.P. Chavan, Customizable object detection using Smartphone In: International Journal of Advanced Research in Computer and Communication Engineering, 2015, 1-4, Vol. 4, Issue 2, *https://ijarcce.com/wp-content/uploads/2015/03/IJARCCE6A.pdf*.</bib>

<bib id="bib52"><number>[52]</number>Varsha Sharma, Chaitanya Sharma, Sahil Jain, Siddhant Jain, Assitance application for visually impaired - vision In: International Journal of Scientific Research and Engineering Development, 2019, 1-4, Volume 2, Issue 6, *http://www.ijsred.com/volume2/issue6/IJSRED-V2I6P52.pdf*.</bib>
<bib id="bib53"><number>[53]</number>Wearable technology: smart glasses and head mounted cameras, Royal National Institute of Blind People, 13 12 2022, *https://www.rnib.org.uk/living-with-sight-loss/assistive-aids-and-technology/tech-support-and-information/wearable-technology-smart-glasses-and-head-mounted-cameras/*.</bib>
<bib id="bib54"><number>[54]</number>Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.</bib>