Social Science and Human Research Bulletin

ISSN(e): 3050-5542

ISSN(p): 3050-5534

Vol. 03(09): 1150-1158, September 2026

Home Page: http://sshrbjournal.org/

From Accessibility Needs to AI System Design: A User-Centered Study of Image-To-Speech Assistive Technology for People with Visual Impairments in Thanh Hoa, Vietnam

Hoang Anh Cong
Hong Duc University, Vietnam

Article DOI: 10.55677/SSHRB/2026-3050-0903 **DOI URL: https://doi.org/10.55677/SSHRB/2026-3050-0903**

KEYWORDS: visual impairment, artificial intelligence, assistive technology, image-to-speech, accessibility, user-centered design, Thanh Hoa.

Corresponding Author
Hoang Anh Cong

Published: September 09, 2026

ABSTRACT: Advances in artificial intelligence (AI), computer vision, optical character recognition (OCR), and speech synthesis are creating new possibilities for helping people with visual impairments access visual information. However, the effectiveness of assistive technology depends not only on model accuracy but also on its fit with users' needs, digital capabilities, and contexts of use. This study identifies information-access barriers, technology-use needs, and functional requirements for a Vietnamese image-to-speech system designed to support people with visual impairments in Thanh Hoa, Vietnam. A mixed-methods descriptive and applied design was employed, combining questionnaire surveys, direct interviews, field observation, and documented field evidence. A total of 230 valid questionnaires were analyzed, comprising 100 responses from people with visual impairments, 120 from guardians/support persons, and 10 from representatives of supporting agencies and organizations. Among the 100 users with visual impairments, 79% had severe visual impairment or complete blindness. The findings indicate particularly strong needs for reading text from images, describing natural scenes, recognizing obstacles and signs, and receiving results in Vietnamese speech. Smartphones provide a promising deployment platform, but users' operational capabilities are uneven, especially in camera use, touch interaction, and error handling. On this basis, the study proposes an architecture integrating image acquisition and quality checking, OCR/AI-based image description, Vietnamese language processing, text-to-speech synthesis, and an audio-first interface. Cross-cutting requirements include minimal interaction steps, audio-guided camera operation, TalkBack/VoiceOver compatibility, rapid feedback, partial offline functionality, and user-data protection. The study demonstrates that translating context-specific user needs into technical requirements is essential for developing AI assistive technologies that are usable in practice and responsive to local conditions.

I. INTRODUCTION

People with visual impairments face persistent barriers to accessing information, education, employment, mobility, and social participation. Whereas sighted users can quickly obtain information from printed text, images, colors, signs, and spatial cues, people with visual impairments often rely on hearing, touch, spatial memory, assistive tools, or support from others. These barriers become more consequential in increasingly digital environments, where documents, public services, educational activities, transactions, and many everyday information sources are organized around visual interfaces. The World Health Organization emphasizes that vision impairment affects not only visual function but also broader dimensions of quality of life, education, employment, and social participation (World Health Organization, 2019).

Assistive technologies can help narrow this accessibility gap. Screen readers such as TalkBack and VoiceOver enable people with visual impairments to use smartphones and access digital content; optical character recognition (OCR) converts text contained in images into machine-readable text; text-to-speech (TTS) transforms text into audio; and contemporary vision-language models can identify objects and generate descriptions of visual scenes. Applications such as Seeing AI and Be My Eyes further illustrate the potential of combining mobile cameras, AI, and audio feedback to support blind and low-vision users. In the broader field of digital preservation in Vietnam, AI and smart technologies have been associated with significant opportunities but also with constraints relating to infrastructure, resources, ethics, and implementation capacity, highlighting the importance of context-sensitive investment and support mechanisms (Phuong, 2026). Similar contextual constraints are relevant to assistive technology. Existing tools do not automatically solve accessibility problems across all countries and user groups because their usefulness depends on language, device infrastructure, smartphone proficiency, connectivity, environmental conditions, and the concrete needs of users.

Image-description technologies have evolved rapidly from early neural captioning architectures such as Show and Tell to increasingly powerful approaches integrating visual and linguistic representations (Vinyals et al., 2015; Herdade et al., 2019). Datasets such as MS COCO have provided important foundations for developing and evaluating image-captioning models (Lin et al., 2014; Chen et al., 2015). For assistive applications, VizWiz is especially relevant because it demonstrates that photographs captured by people with visual impairments in real-world conditions may differ substantially from conventional benchmark images: they can be tilted, blurred, poorly illuminated, partially cropped, or difficult to frame around the intended object (Gurari et al., 2020). This observation has direct implications for image-to-speech systems because the quality of image input is closely connected to users' ability to operate a camera without relying on vision.

Field evidence from Thanh Hoa indicates that people with visual impairments encounter diverse difficulties when reading printed materials, administrative documents, medicine and product labels, identifying signs, detecting obstacles, understanding surrounding environments, and operating smartphones. Users therefore need more than a text-reading tool; they also need assistance in accessing non-textual visual information relevant to everyday activities and safety. The recurring functions identified in the field data include reading text from images, describing scenes, recognizing obstacles, identifying signs, and delivering results through Vietnamese speech.

Assistive technology should therefore not be developed from a purely technical perspective. An AI model may achieve high recognition accuracy and still remain unsuitable if the user must perform complex camera alignment, visually inspect substantial on-screen information, or navigate too many interaction steps. Conversely, evidence about user needs can be translated directly into system and interface requirements: difficulty reading printed text implies Vietnamese OCR and TTS; difficulty understanding surroundings implies image description; difficulty operating a camera implies audio-guided image capture; and difficulty with touchscreens implies simplified interfaces and voice interaction.

Against this background, the study has three objectives. First, it assesses barriers related to information access, technology use, and environmental interaction among people with visual impairments in Thanh Hoa. Second, it identifies functional needs and accessibility requirements for an AI-based Vietnamese image-to-speech system. Third, it translates user-derived findings into system architecture and design principles for an assistive application. The main contribution lies in connecting field-based evidence with software requirements engineering rather than treating user research and technology development as separate processes.

II. THEORETICAL BACKGROUND AND RESEARCH METHODS

2.1. Assistive technology and information accessibility for people with visual impairments

Assistive technology is intended to maintain or improve users' functioning, independence, and social participation. For people with visual impairments, relevant technologies include tools that translate visual information into auditory or tactile forms, screen readers, OCR systems, object-recognition and image-description technologies, mobility and orientation aids, and voice-based interactive assistants.

Within digital environments, TalkBack on Android and VoiceOver on iOS provide an important accessibility layer by allowing users to listen to interface content and interact with applications. Nevertheless, screen readers mainly facilitate access to information that already exists in digital form. They do not automatically address situations such as reading a printed sheet, identifying the content of a medicine label, recognizing a sign in the physical environment, or understanding a surrounding scene. The Thanh Hoa survey also indicates that many users have some experience with Android/TalkBack or iPhone/VoiceOver, but proficiency is uneven and camera operation remains a substantial barrier.

An image-based assistive system therefore requires the integration of several components rather than reliance on a single technology. When an image contains text, OCR can extract textual content; when it contains a natural scene, a vision-language model can generate a description; natural language processing can then normalize the output in Vietnamese before TTS converts it into spoken information. This integration forms an image-to-Vietnamese-information-to-Vietnamese-speech processing chain.

2.2. User-centered design in assistive technology

For technologies designed for people with disabilities, accessibility should be treated as a core product attribute rather than as an additional feature introduced after software development. The United Nations Convention on the Rights of Persons with Disabilities emphasizes the right of persons with disabilities to access information, communications, and technologies (United Nations, 2006). In digital environments, WCAG 2.2 likewise emphasizes perceivability, operability, and access through multiple modes of interaction (W3C, 2023).

For image-to-speech software, user-centered design means that technical decisions should be grounded in actual contexts of use. If a user cannot determine the direction of the camera, for example, simply improving recognition accuracy does not solve the accessibility problem. The application should actively guide camera positioning, indicate when an image is poorly illuminated or blurred, and provide audio or haptic feedback. Similarly, during mobility, a concise warning about an obstacle or an open route may be more useful than a long and detailed scene description.

The study therefore follows a translation logic from barrier to need, from need to functional requirement, from functional requirement to technical component, and from technical component to implementation criterion. This logic is consistent with requirements engineering, in which system requirements should be derived from stakeholder needs and real conditions of use. In a different application domain in Vietnam, Phuong and Lam (2025) found that local-community feedback made a meaningful contribution to heritage-communication effectiveness, underscoring the value of direct participation and contextual adaptation. Although the substantive field is different, this evidence supports the methodological principle that end users of assistive technology should participate in requirements identification and solution evaluation.

2.3. Research design

The study employs a mixed-methods descriptive and applied research design. Data were collected through four complementary sources: questionnaire surveys, direct interviews, field observation, and documentary evidence including photographs, videos, audio recordings, and field logs. Following collection, questionnaires were checked, classified, coded, and entered for analysis. Quantitative data were summarized using frequencies and percentages, whereas qualitative material from interviews and observations was organized thematically.

The field survey was implemented during the needs-and-current-status assessment phase of the broader research project, from October 2025 to March 2026. The multi-source design enabled the research team to compare self-reported information with observations and field evidence and to use qualitative findings to interpret the descriptive quantitative results.

2.4. Participants and survey scale

The study involved three groups of information providers: (i) people with visual impairments, who constituted the direct end-user group; (ii) guardians/support persons, who regularly observed and assisted people with visual impairments; and (iii) representatives of agencies and organizations supporting people with visual impairments, who provided information on facilities, technological infrastructure, and implementation readiness.

A total of 230 questionnaires were collected, and all 230 were retained as valid for analysis after data checking. These comprised 100 questionnaires from people with visual impairments, 120 from guardians/support persons, and 10 from agencies or organizations.

Data were collected across different use environments, including Blind Associations, centers/schools/support facilities, and community or household settings. Specifically, 90 questionnaires were collected through Blind Association settings, 70 through centers, schools, or other support facilities, and 70 in community or household contexts. This distribution allowed the study to capture both institutionally supported use conditions and everyday individual-use conditions.

Table 1. Structure of the survey sample

Respondent group	Number	Percentage (%)	Role in the survey
People with visual impairments	100	43.48	Provide direct information on needs, barriers, and technology-use capabilities
Guardians/support persons	120	52.17	Provide complementary information on daily living, learning, mobility, and technology support
Agencies/organizations	10	4.35	Assess facilities, technological infrastructure, and implementation readiness
Total	230	100.00	

Source: *Compiled from the project survey data.*

2.5. Data collection content

For people with visual impairments, the questionnaire focused on severity of visual impairment; use of devices, smartphones, and accessibility tools; difficulties in reading text and recognizing objects, obstacles, signs, and environments; and needs relating to AI-assisted software.

For guardians/support persons, the survey examined the intensity and types of support provided, situations in which people with visual impairments required assistance, technology-use capabilities, and support persons' assessments of functions that should be prioritized.

For agencies and organizations, the survey covered facilities, Internet connectivity, devices, information-technology support personnel, capacity to organize training, and willingness to participate in software trials.

In addition to questionnaires, direct interviews were used when participants were unable or uncomfortable reading or completing forms independently. Field observation documented living and learning spaces, devices, network infrastructure, and how people with visual impairments interacted with smartphones. Photographs, videos, audio recordings, and field logs were used to support data checking and triangulation.

2.6. Data processing and confidentiality principles

Quantitative data were summarized by counts and percentages. Qualitative data were grouped around themes including text access, environmental recognition, mobility, digital devices, camera operation, voice interaction, interface requirements, and implementation conditions.

Survey information was coded by respondent group. When images, audio, or personal information were collected, participants were informed of the purpose of the survey and the voluntary nature of participation; audiovisual evidence was collected on the basis of consent. Potentially sensitive information was to be managed for research purposes and its use limited to what was necessary for the study.

III. RESEARCH FINDINGS

3.1. Severity of visual impairment among end users

Among the 100 participants with visual impairments, 21 had mild visual impairment, 38 had severe visual impairment, and 41 were completely blind. Thus, the combined severe-impairment and blindness groups accounted for 79% of direct end users. This distribution is important because most potential users cannot rely on visual interfaces in the same way as sighted users.

Table 2. Severity of visual impairment among participating end users

Level of visual impairment	Number	Percentage (%)
Mild visual impairment	21	21.0
Severe visual impairment	38	38.0
Complete blindness	41	41.0
Total	100	100.0

Source: *Compiled from the project survey data.*

This structure indicates that a system requiring visual inspection of the screen or visually guided camera operations would be poorly suited to the main user group. Audio feedback, step-by-step guidance, and the ability to operate the application without looking at the screen must therefore be treated as essential rather than optional design requirements.

3.2. Barriers to accessing visual information

The survey indicates that information-access barriers are not confined to a single type of content. In daily living, users need to read medicine labels, food labels, receipts, and instructions. In education, needs include books, textbooks, photocopied materials, exercises, and announcements. At work and in public services, difficulties arise with forms, contracts, appointment documents, directions, and administrative notices. In community settings, shop signs, house numbers, room signs, and public notice boards create additional recurring barriers.

These situations explain why Vietnamese OCR and TTS are core functions. Yet recognizing characters alone is insufficient. Users need the system to read clearly, support replay and speed adjustment, and handle non-ideal images such as tilted, blurred, or poorly illuminated photographs.

3.3. Need for scene description and environmental recognition

Beyond written text, many difficult situations involve non-textual visual information. Users need to know what is in front of them, where objects are located, whether a route is blocked, where a sign or shop is positioned, and whether environmental elements may create a safety risk.

The need for scene description appears across multiple contexts: identifying objects and household items; accessing images or diagrams in learning; recognizing pathways and vehicles during mobility; locating signs in community settings; and recognizing equipment or spatial arrangements in the workplace.

An important finding is that the content of descriptions should adapt to the context. In learning situations, descriptions may be more detailed; during mobility, the system should prioritize safety-relevant information such as obstacles, steps, vehicles, and route conditions. Lengthy descriptions in situations requiring rapid response may increase the user's information-processing burden rather than improve accessibility.

3.4. Need for obstacle recognition and mobility support

Mobility difficulties center on detecting obstacles, distinguishing sidewalks from roadways, locating entrances and exits, recognizing signs, and understanding unfamiliar environments. Obstacles were reported across settings, ranging from furniture and steps indoors to motorcycles, poles, advertising boards, and goods encroaching on pathways in public spaces.

Users need more than a binary notification that an obstacle is present. They also need information about its location and relevance. Accordingly, messages such as “there is a step ahead” or “there is a motorcycle on the left” are potentially more useful in practice than a long description of the entire scene.

AI should nevertheless not be positioned as a replacement for a white cane, orientation skills, or other established mobility supports. Especially in dense traffic or high-risk environments, an image-to-speech system should function as an additional source of information to support, rather than determine, user decisions.

3.5. Smartphone usability and the gap between device access and effective use

Smartphones constitute a suitable deployment platform because they integrate cameras, microphones, speakers, network connectivity, and screen-reading tools. The survey shows that many people with visual impairments already use Android devices or iPhones, and some are familiar with TalkBack or VoiceOver. This creates a favorable initial foundation for deployment. However, proficiency is uneven.

The most notable difficulty for an image-to-speech application concerns camera operation. Users may not know where the camera is pointing, whether the intended object is within the frame, or whether the image is blurred or underexposed. Consequently, an AI system can perform well at the model level and still fail in use if it does not assist users in producing usable image input.

The findings suggest several user groups: proficient smartphone users, basic users, users with limited smartphone skills, and users who rely on feature phones or do not own a smartphone. Each group requires a different deployment strategy, ranging from independent early trials to direct guidance or access through shared devices at supporting organizations.

3.6. Priority needs and functional system requirements

Synthesis of the survey results indicates that the needs most directly related to system design can be translated into corresponding functions and technical components, as summarized in Table 3.

Table 3. Translating user needs into technology functions

User need/barrier	Required function	Technical component
Difficulty reading documents, forms, and product labels	Read text from images	Vietnamese OCR + TTS
Difficulty identifying scenes and objects	Describe images/scenes	Vision-language model + NLP
Difficulty recognizing obstacles	Detect and warn about safety-relevant elements	Object/scene recognition + TTS
Difficulty operating the camera	Provide capture guidance and image-quality checking	Image-quality analysis + audio/haptic cues
Difficulty using touch controls	Provide voice control and a simplified interface	Voice interaction + accessibility APIs

User need/barrier	Required function	Technical component
Unstable network connectivity	Maintain basic offline functions	On-device processing + local storage
Cannot visually monitor system status	Provide audio status feedback	Audio feedback

Source: *Author's proposal based on the survey findings.*

Across the evidence, reading text from images, scene description, obstacle recognition, and spoken Vietnamese output emerged as the highest-priority needs. Mobility support, voice control, and support for learning and work were also identified as necessary functional groups.

3.7. Interface and interaction requirements

The findings indicate that the interface should follow a voice-first principle rather than simply add accessibility functions to an interface originally designed for sighted users.

Core functions should be limited to clear actions such as “take a photo,” “read text,” “describe scene,” “replay,” and “exit.” All controls should be properly labeled so that TalkBack and VoiceOver can identify them. When the application opens, it should provide brief audio guidance instead of requiring users to read instructions from the screen.

Voice control is also important. Short commands such as “take a photo,” “read text,” “describe scene,” “replay,” or “stop reading” reduce the need for touch interaction. After each command, the application should confirm receipt of the request and communicate the current processing status.

3.8. Response speed and feedback architecture

Response speed has different implications for different tasks. For long documents, users may accept slightly longer processing times in exchange for higher accuracy. In contrast, for obstacle recognition or rapid interpretation of the scene ahead, latency should be low and the output concise.

A layered feedback architecture is therefore appropriate. The system can first provide the most important information, for example, “an obstacle is ahead,” and provide a more detailed description only if the user requests it. The application should also communicate status messages such as “processing image” and warn users when the image is unclear or network connectivity is weak.

3.9. Offline functionality requirements

Variation in network conditions across locations and users means that the system should not depend entirely on online processing. Lightweight functions such as opening the application, providing operational guidance, replaying stored results, and communicating system status can be performed on the device. More computationally demanding tasks, such as detailed scene description or large-model image analysis, may still require server-side processing.

A hybrid on-device/online architecture can therefore balance speed, quality, and availability. When connectivity is unavailable, the application should communicate the problem through speech rather than relying only on a visual error icon.

3.10. Proposed image-to-speech system architecture

Based on the user survey and the technical design, the system can be organized as the following processing chain: image from camera/library -> image-quality checking and preprocessing -> OCR or AI-based image analysis/description -> Vietnamese-language normalization -> text-to-speech conversion -> audio output to the user.

The technical system has been designed around use cases including AI-based image processing, automatic caption generation, result evaluation, and text-to-speech conversion. Once a Vietnamese description is produced, the TTS component generates an audio file that can be played directly by the mobile application.

The prototype has also established user functions for selecting or capturing an image, converting image content into speech, saving processing history, and replaying previous results. A supporting data package organizes 30,000 images into 26,153 training images, 1,529 validation images, and 2,318 test images. The data structure links images, annotations, audio, and metadata for subsequent training and evaluation.

IV. DISCUSSION

The findings indicate that the central challenge in developing image-to-speech technology for people with visual impairments is not simply how accurately an AI model recognizes an image, but whether the entire interaction chain is usable by the intended users. This distinction separates model-centered AI development from assistive-technology-centered AI design.

First, OCR remains a foundational function. Printed text appears in almost every area of everyday life, including medicine labels, learning materials, administrative documents, directions, and product information. A tool that can rapidly convert such

information into spoken Vietnamese can reduce reliance on another person to read it aloud. However, the findings also expose the limits of an OCR-only approach. Everyday life also contains substantial non-textual information, including object positions, obstacles, scenes, and spatial relationships. Image description should therefore be treated as a necessary complement to OCR rather than as an optional extension.

Second, the survey demonstrates that input quality is an accessibility issue, not only a technical issue. Conventional machine-learning datasets often contain relatively clear and intentionally framed images. For people with visual impairments, image capture can fail before inference begins because the user cannot visually monitor the camera. This observation is consistent with experience from assistive datasets such as VizWiz (Gurari et al., 2020). Camera guidance should therefore be treated as part of the AI pipeline itself: orientation guidance, blur and illumination checking, and retake prompts should occur before model inference.

Third, feedback design should be context dependent. In conventional image-captioning tasks, more complete descriptions may sometimes be treated as better descriptions. In mobility support, this is not always the case. A user who is walking may need to hear “there is a step ahead” immediately rather than listen to a multi-sentence account of the whole scene. An assistive image-to-speech system should therefore prioritize information according to relevance and safety, not only linguistic completeness.

Fourth, the fact that 79% of direct end users in the sample had severe visual impairment or complete blindness has significant design implications. An interface in which accessibility is added at the end of development is unlikely to meet the needs of this user group. Audio output, voice interaction, accessible labels, haptic feedback, and screen-reader compatibility need to be embedded in the interaction architecture from the beginning.

Fifth, the findings show the need to distinguish between smartphone ownership and the ability to use a smartphone effectively for camera-based AI tasks. Users may be competent at calls, messaging, or TalkBack while still experiencing major difficulty orienting the camera. Smartphone penetration alone should therefore not be used as evidence that an assistive AI solution is practically deployable.

Sixth, the findings suggest a tiered implementation model. Users with strong smartphone skills can participate in early trials; basic users require direct training; users with limited skills need support from relatives or Blind Association staff; and users without smartphones may initially access the system through shared devices at supporting institutions. This is more realistic than assuming that every user will immediately install and operate the application independently. In the context of digital preservation in Vietnam, the implementation of smart technologies has likewise been shown to face uneven infrastructure, financial resources, technical capacity, and ethical requirements across locations and institutions (Phuong, 2026). Although the application domain differs, these constraints provide a useful contextual reference for locally deployed assistive technology.

Finally, AI should be positioned within an ecosystem of assistive supports. An image-to-speech application does not replace TalkBack/VoiceOver, white canes, guides, or personal orientation skills; rather, it adds types of visual information that these supports do not fully provide. This complementary positioning also reduces the risk of overstating AI capabilities, especially in situations involving physical safety.

V. CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

This study used 230 valid questionnaires from three respondent groups - people with visual impairments, guardians/support persons, and supporting agencies/organizations - to identify information-access barriers and translate them into development requirements for an AI-based image-to-speech system. The structure of the sample enabled the problem to be examined from the perspectives of direct users, those who support them, and the environments in which implementation may occur.

The findings show that the needs of people with visual impairments extend beyond text reading. Important functions include reading text from images, describing scenes, recognizing obstacles and signs, supporting image capture, and providing Vietnamese voice feedback. Among the 100 direct users surveyed, 79% had severe visual impairment or complete blindness, indicating that the system must be operable without reliance on visual monitoring.

Smartphones provide a feasible deployment platform because they combine cameras, microphones, speakers, connectivity, and accessibility tools. Nevertheless, users' operational skills vary, and camera operation is a major bottleneck. Audio guidance, image-quality checking, voice control, and TalkBack/VoiceOver compatibility should therefore be treated as core requirements.

From a technical perspective, an appropriate architecture is not a single AI model but an integrated processing chain that combines image acquisition, image-quality checking, OCR/image description, Vietnamese language processing, TTS, and audio feedback. The architecture must balance accuracy and speed with accessibility and the actual device and network conditions of users.

5.2. Practical recommendations

First, software development should continue through co-design with people with visual impairments. Trial versions should be evaluated by users with different levels of vision and smartphone skills, rather than being assessed only through technical testing by developers.

Second, the initial trial should prioritize a minimum set of high-value functions: reading text from images, concise scene description, camera guidance, Vietnamese spoken output, and replay. Introducing too many functions at an early stage may increase interface complexity.

Third, evaluation should be context specific. OCR functions should be assessed primarily for recognition accuracy and readable output; scene description should be assessed for relevance and usefulness; and obstacle alerts should prioritize response speed and safety-related utility.

Fourth, the application should adopt a voice-first design with complete accessibility labels, audio/haptic feedback, short voice commands, and a low-step interaction flow. Compatibility with TalkBack and VoiceOver should be evaluated as an independent acceptance criterion.

Fifth, a hybrid online-offline architecture should be implemented. Interface navigation, guidance, history access, and status feedback should remain available without network connectivity, whereas computationally intensive AI tasks can be processed on a server when required.

Sixth, image, text, and speech data should be governed through clear provenance, quality control, and personal-data protection. Dataset development should be treated as a continuing process informed by real use cases rather than as a one-time collection activity (Gebru et al., 2021; Whang et al., 2023).

Seventh, trials should begin at Blind Associations, centers, or other facilities capable of supporting users and should then be expanded to community and household settings. A phased approach can reduce implementation risk while creating a systematic feedback mechanism for software improvement.

5.3. Limitations and directions for future research

The study has several limitations. First, data were collected from people with visual impairments and related organizations at the project's survey sites; the descriptive percentages should therefore not be mechanically generalized to the entire population of people with visual impairments in Thanh Hoa.

Second, the present findings focus primarily on needs, conditions of use, and design requirements. They do not provide sufficient evidence to conclude that the final version of the software improves independence or safety. Such outcomes require user testing after the system has reached a sufficiently stable stage of development.

Third, technical indicators including OCR accuracy, scene-description quality, latency under real operating conditions, speech intelligibility, and independent task-completion rates need to be measured through controlled evaluation designs.

Future research should conduct usability trials with end users and measure task-completion time, interaction errors, level of assistance required, output accuracy, satisfaction, and intention to continue using the system. Comparative analysis should also examine differences by severity of visual impairment, age, and smartphone proficiency. Functions related to mobility and obstacle detection should first be tested in controlled and safe environments before evaluation in public settings.

REFERENCES

1. Apple. (n.d.). Turn on and practice VoiceOver on iPhone. Apple Support. https://support.apple.com/guide/iphone/turn-on-and-practice-voiceover-iph3e2e415f/ios
2. Be My Eyes. (n.d.). Be My Eyes app: Bringing sight to blind and low-vision people. https://www.bemyeyes.com/
3. Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. arXiv:1504.00325. https://arxiv.org/abs/1504.00325
4. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3458723
5. Google. (n.d.). Get started on Android with TalkBack. Android Accessibility Help. https://support.google.com/accessibility/android/answer/6283677
6. Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., & Bigham, J. P. (2020). VizWiz: Nearly real-time answers to visual questions. International Journal of Computer Vision, 128, 1-20. https://doi.org/10.1007/s11263-019-01266-6
7. Herdade, S., Kappeler, A., Boakye, K., & Soares, J. (2019). Image captioning: Transforming objects into words. Advances in Neural Information Processing Systems, 32, 11137-11147.
8. Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3128-3137. https://doi.org/10.1109/CVPR.2015.7298932
9. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In Computer Vision - ECCV 2014 (pp. 740-755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48
10. Microsoft. (2023, December 4). Seeing AI app launches on Android: Including new and updated features and new languages. Microsoft Accessibility Blog.

11. Phuong, N. T. (2026). Digital preservation of Vietnamese cultural heritage: Opportunities and limitations in the age of smart technology. Preservation, Digital Technology & Culture, 55(2), 117-129. https://doi.org/10.1515/pdtc-2025-0035
12. Phuong, N. T., & Lam, N. T. N. (2025). Evaluating the effectiveness of cultural heritage communication based on local community feedback: A case study of Hanoi city. Scientific Culture, 11(3), 1-11. https://doi.org/10.5281/zenodo.17379320
13. Smith, R. (2007). An overview of the Tesseract OCR engine. Ninth International Conference on Document Analysis and Recognition, 629-633. https://doi.org/10.1109/ICDAR.2007.4376991
14. United Nations. (2006). Convention on the Rights of Persons with Disabilities. https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html
15. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3156-3164. https://doi.org/10.1109/CVPR.2015.7298935
16. W3C. (2023). Web Content Accessibility Guidelines (WCAG) 2.2. World Wide Web Consortium. https://www.w3.org/TR/WCAG22/
17. Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. Interspeech 2017, 4006-4010. https://doi.org/10.21437/Interspeech.2017-1452
18. Whang, S. E., Roh, Y., Song, H., & Lee, J.-G. (2023). Data collection and quality challenges in deep learning: A data-centric AI perspective. The VLDB Journal, 32, 791-813. https://doi.org/10.1007/s00778-022-00775-9
19. World Health Organization. (2019). World report on vision. World Health Organization. https://www.who.int/publications/i/item/9789241516570
20. World Health Organization. (2023). Disability and health. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/disability-and-health