Improving Navigation Systems with Computer Vision Approaches

Artem Panchenko[1], Denys Drobin[2], Kyrylo Rukkas[3], Anastasiia Morozova[4], Lyudmyla Polyakova[5] and Iryna Zaretska[6]

V.N. Karazin Kharkiv National University, 4 Svobody sq., Kharkiv, 61101, Ukraine

Abstract

This work presents an analysis of existing assistive technologies for people with visual impairments and introduces a specialized mobile application designed to ensure safe navigation in indoor environments. The study provides an overview of computer vision methods and semantic segmentation techniques applied to navigation tasks for people with disabilities. To estimate safe walking distances, a real-time depth estimation algorithm was proposed. The developed methods were integrated into a mobile application and tested under real-world indoor conditions. The accuracy of the system was evaluated through experiments conducted across various locations.

Keywords

computer vision, navigation, software development, neural networks, semantic segmentation, image processing, image segmentation model, database

1. Introduction

Vision, as the most dominant human sense, is essential for virtually all aspects of daily life. While often taken for granted, the absence or impairment of vision significantly hinders the ability to learn, move independently, read, and participate fully in educational and professional activities. Visual impairment arises when an ocular condition disrupts the normal functioning of the visual system. It is estimated that, over a lifetime, nearly every individual will experience at least one eye condition requiring appropriate medical attention.

According to research conducted by the World Health Organization [1], approximately 2.2 billion people worldwide experience various forms of visual impairment. Among them, roughly 1.1 billion individuals have conditions that affect their ability to see clearly and distinguish objects at a distance. Notably, around 200 million people suffer from conditions that prevent them from navigating safely and independently.

To facilitate spatial orientation, individuals with visual impairments typically rely on the use of a white cane [2]. This tool enables users to detect obstacles along their path, such as curbs, stairs, or objects on the floor, and also signals to others that the user has a visual impairment. However, the white cane has several limitations that can significantly affect both safety and overall user experience:

1. Limited range of detection – the cane can only detect objects within immediate proximity, which restricts the user's ability to plan longer routes.
2. Incomplete detection of obstacles – low-lying or elevated obstacles, such as objects at head level or narrow protrusions, may go unnoticed, potentially compromising safety.
3. Reduced effectiveness in complex or crowded environments – narrow corridors, stairways with obstacles, and densely populated areas can hinder safe and efficient use of the cane.
4. Dependence on user skill – effective navigation with the cane requires specialized training; without it, users may struggle to orient themselves efficiently.

ProfIT AI'25: 5th International Workshop of IT-professionals on Artificial Intelligence, October 15–17, 2025, Liverpool, UK

artem.panchenko@karazin.ua (A. Panchenko); drobin2022mf11@student.karazin.ua (D. Drobin); rukkas@karazin.ua (K. Rukkas); a.morozova@karazin.ua (A. Morozova); l.yu.polyakova@karazin.ua (L. Polyakova); zaretskaya@karazin.ua (I. Zaretska)

0000-0001-5865-6158 (A. Panchenko); 0000-0002-7614-0793 (K. Rukkas); 0000-0003-2143-7992 (A. Morozova); 0000-0002-6674-1958 (L. Polyakova); 0000-0001-8747-2737 (I. Zaretska)

The present study focuses on addressing the aforementioned limitations associated with the use of the white cane. We propose the development of a mobile application leveraging Computer Vision technologies [3], which aims to enhance the mobility and safety of individuals with visual impairments in the following ways:

1. The use of a smartphone camera for navigation removes the distance limitations inherent to a white cane and enables the detection and avoidance of all obstacles, rather than only those located on the floor.
2. A mobile application also facilitates efficient navigation in confined spaces, as it eliminates the need for active manipulation of a cane with the hand.
3. Moreover, the challenge of learning to use a white cane—which typically requires specialized facilities and professional supervision—is addressed. By leveraging a mobile application for spatial orientation, visually impaired individuals can independently develop navigation skills while maintaining a high level of safety, with the smartphone providing real-time information about obstacles along the route.

The foregoing considerations underscore the primary objective of this study: the design and development of a mobile application specifically aimed at assisting individuals with visual impairments in navigating indoor environments safely and efficiently. Enclosed spaces, including shopping centers, office buildings, hospitals, and residential complexes, present a variety of potential hazards for untrained individuals moving without adequate visual guidance. These hazards may include obstacles at different heights, narrow passageways, staircases, and dynamic elements such as other people or moving objects. In such contexts, traditional assistive tools, such as the white cane, provide only limited coverage and require specialized training to use effectively. The proposed mobile application seeks to address these limitations by integrating computer vision technologies to provide real-time information about the surrounding environment, detect obstacles of varying sizes and positions, and guide users along safe paths. By doing so, this solution aims to substantially improve both the autonomy and the safety of visually impaired individuals in complex indoor settings, offering a practical complement—or in some cases an alternative—to conventional mobility aids.

This paper should be regarded as both an introduction to the research area and a proof of concept for a mobile application designed to facilitate navigation for individuals with visual impairments. The paper is structured as follows:

- Section 3 presents the advantages and methodological approaches for the practical implementation of a contemporary solution to computer vision tasks, specifically focusing on Image Semantic Segmentation.
- Section 4 details the architectural design and software engineering decisions underlying the developed application.
- Section 5 demonstrates the application's usage in practice, illustrating its operational workflow, user interaction, and functional capabilities.

2. Related Work

2.1. Adaptation of Mobile Devices for Visually Impaired Users

Modern smartphones are equipped with a variety of accessibility features by default, among which screen readers play a central role. These tools provide an audio-based interface for interacting with the device and are comprehensive solutions offered by both iOS and Android operating systems. Screen readers operate by converting visual information into audio output, allowing users to navigate the interface through a combination of voice feedback and gesture-based commands.

According to a study conducted by WebAIM, which surveyed nearly 2,000 visually impaired individuals, of whom 76.6% were completely blind, almost 90% of respondents rely on smartphones equipped with screen readers as their primary means of communication and information access [4].

Despite their effectiveness, current screen reader systems have notable limitations. They are primarily designed for digital content navigation and are less effective for real-world spatial awareness or obstacle detection. Users must rely on traditional mobility aids, such as the white cane, for safe navigation in unfamiliar environments.

2.2. Be My Eyes

Be My Eyes is a platform that leverages real-time video communication to connect visually impaired users with a network of sighted volunteers [5]. The application assists users in reading text, navigating environments, and identifying objects by having volunteers provide audio descriptions of the visual input captured through the user's smartphone camera.

A key advantage of this approach lies in its high accuracy and contextual understanding, as human volunteers can interpret subtle details and nuances that automated systems may miss. This allows for more precise and situation-specific guidance compared to purely algorithmic solutions.

However, the platform also exhibits notable limitations. Its functionality depends on the availability of volunteers and requires a constant Internet connection, which may restrict usability in areas with limited connectivity or during periods of high demand. Additionally, reliance on human volunteers can introduce variability in response time and guidance quality, which may affect the overall reliability of the service.

2.3. Seeing AI (Microsoft)

Seeing AI by Microsoft is a standalone application that employs artificial intelligence algorithms to analyze the surrounding environment using a smartphone camera [6]. The application provides a wide range of object and image recognition functions, including short text and document recognition, product identification via barcodes, detection of major currency denominations, and more. These capabilities offer substantial benefits to visually impaired individuals, facilitating daily tasks and enhancing their autonomy and independence.

A key advantage of Seeing AI is its ability to operate in real time even with limited network access, thanks to its neural network architecture optimized for local computations. Unlike human-assisted systems such as Be My Eyes, Seeing AI functions independently of volunteers, ensuring user autonomy and complete privacy.

Despite these strengths, the system also presents notable limitations. Automated AI-based solutions can struggle with complex contextual scenes, particularly in environments with intricate spatial arrangements or unusual lighting conditions. Additionally, while the system provides consistent performance, it may lack the nuanced contextual understanding that a human volunteer could provide in ambiguous or unexpected situations.

2.4. Conclusions

In summary, it can be observed that individuals with visual impairments widely utilize a variety of mobile applications to enhance their quality of life, particularly in the areas of navigation and as a supplement or alternative to the traditional white cane. Notably, there is an increasing adoption of artificial intelligence techniques, especially computer vision, to address the challenges associated with safe and autonomous mobility for visually impaired users. These technologies enable real-time environmental analysis, obstacle detection, and contextual understanding, providing users with enhanced independence and confidence in navigating both familiar and unfamiliar spaces.

An analysis of the most popular applications on the market, such as Seeing AI and Be My Eyes, allows for the identification of several key requirements for developing effective assistive products:

1. Offline Functionality: The application should be capable of operating without a continuous Internet connection. This requirement stems from the fact that users may encounter locations

where stable connectivity is unavailable, such as in subway systems, trains, or remote indoor environments. Offline operation ensures consistent accessibility and safety in these scenarios.

2. Focused Operational Context: It is essential to concentrate the application's functionality on a specific type of environment. Attempting to address too wide a spectrum of use cases can reduce accuracy, primarily due to undertraining of computer vision models. By tailoring the application to well-defined locations, the system can achieve higher precision in obstacle detection, scene understanding, and overall user guidance.

3. Image Semantic Segmentation

In the field of computer vision, several primary directions have emerged in image processing. One of the earliest and simplest tasks is image classification, in which a model identifies the dominant object within an image and provides an overall assessment of the scene. Classification with localization extends this approach by not only identifying the object but also indicating its position within the scene using a bounding box.

A more complex task is multi-object classification, which involves detecting and categorizing multiple objects within a single image. This challenge requires advanced neural network architectures, such as YOLO [7] or R-CNN [8], capable of handling overlapping and spatially complex objects. Additionally, the Mask R-CNN algorithm not only classifies objects but also generates precise segmentation masks for each individual instance. This instance-level segmentation is particularly important in scenarios where objects of the same class overlap, as conventional bounding boxes alone are insufficient for accurate scene understanding.

Semantic segmentation [9] aims to classify each individual pixel in an image, partitioning the scene into meaningful semantic regions. Unlike simpler tasks such as image classification, which assigns a single label to the entire image, or object detection, which localizes objects using bounding boxes, semantic segmentation provides precise identification of objects and their boundaries. By analyzing each pixel and its interaction with surrounding pixels, this task becomes computationally intensive, posing significant challenges for deployment on mobile devices with limited processing power and energy constraints.

Semantic segmentation appears to be the most suitable approach for identifying safe surfaces for navigation for several key reasons. The task requires the detection of continuous safe surfaces, which differs substantially from conventional object-level classification tasks. For visually impaired individuals learning to navigate with a white cane, understanding the holistic structure of traversable surfaces, including their boundaries and potential obstacles, is essential. This pixel-level comprehension enables the creation of safer and more reliable navigation systems that provide real-time guidance in complex indoor environments.

In summary, semantic segmentation represents the most suitable computer vision approach for solving the problem of safe navigation for individuals with visual impairments. Unlike classification or object detection methods, which provide only coarse information about objects or their locations, semantic segmentation delivers a detailed, pixel-level representation of the environment. This enables the detection of continuous traversable surfaces, accurate boundary delineation, and the identification of potential collision hazards. Such capabilities are essential for ensuring safety and reliability in assistive navigation systems. Therefore, semantic segmentation should be regarded as the optimal technological foundation for the development of applications aimed at enhancing independent mobility for visually impaired users.

4. Application Implementation

In this section, the detailed implementation of the proposed mobile application will be presented. Building upon the previously outlined analysis of computer vision techniques, particular emphasis will be placed on integrating semantic segmentation as the core technology for ensuring safe navigation. The

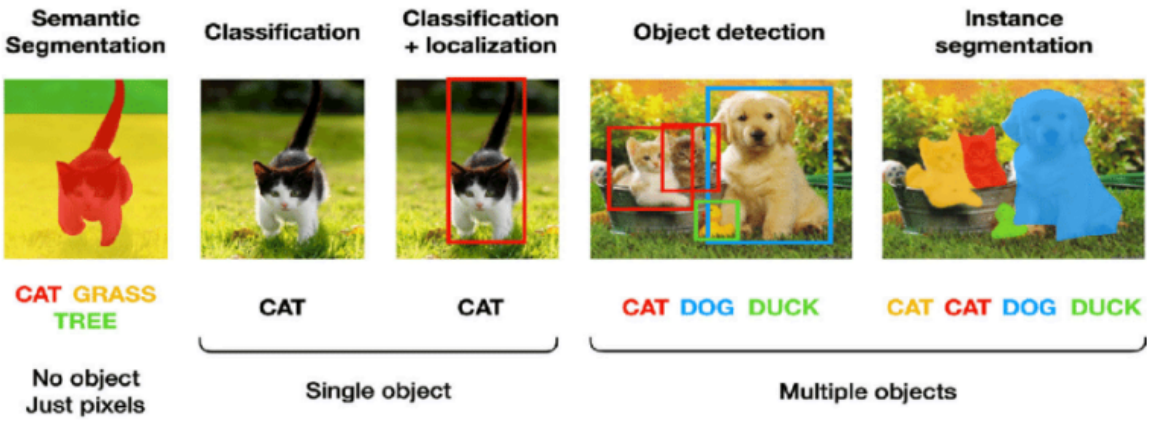

Figure 1: Computer vision task areas

section will describe the system architecture, the choice of neural network models, and the optimization strategies adopted to enable real-time performance on mobile devices with limited computational resources. Additionally, attention will be given to the user interface design and accessibility features, ensuring that the application can be effectively utilized by individuals with visual impairments in real-world conditions.

4.1. System Architecture Description

The developed information system for the identification of safe mobility surfaces employs a multi-layered architecture that ensures efficient interaction among system components and optimal utilization of the computational resources of a mobile device. The system architecture consists of four principal layers: the presentation layer, the business logic layer, the data processing layer, and the machine learning model layer. Each layer is responsible for a specific set of functionalities:

- Presentation Layer. The presentation layer is implemented through the MainActivity component, which is responsible for user interaction and the visualization of analysis results. This component incorporates user interface elements such as buttons for selecting images, capturing photos via the camera, and display areas for both input and processed images. At this level, mechanisms for handling permissions related to camera and storage access are implemented, alongside the interaction logic with Android system components required for acquiring images.
- Business Logic Layer. The business logic layer is represented by the functions processBitmapInternal and processSelectedImage, which receive images from the presentation layer and prepare them for subsequent utilization at the data processing layer. At this stage, asynchronous processing routines are employed, enabling efficient data handling without monopolizing the computational resources of the mobile device.
- Data Processing Layer. The data processing layer comprises image transformation functions such as rotateBitmap, which adjusts the orientation of the image, and bitmapToArray, which converts the raw image representation into a structured data array. At this stage, preliminary image preprocessing is performed to ensure the correct operation of the subsequent neural network algorithms.
- Machine Learning Model Layer. The machine learning model layer is represented by the OnnxSegmentationProcessor component, which encapsulates the interaction logic with the neural network implemented in the ONNX format. This component performs semantic segmentation of images in order to identify safe mobility surfaces. By abstracting the details of model execution, the OnnxSegmentationProcessor ensures modularity and facilitates the integration of the trained deep learning model into the mobile system architecture.

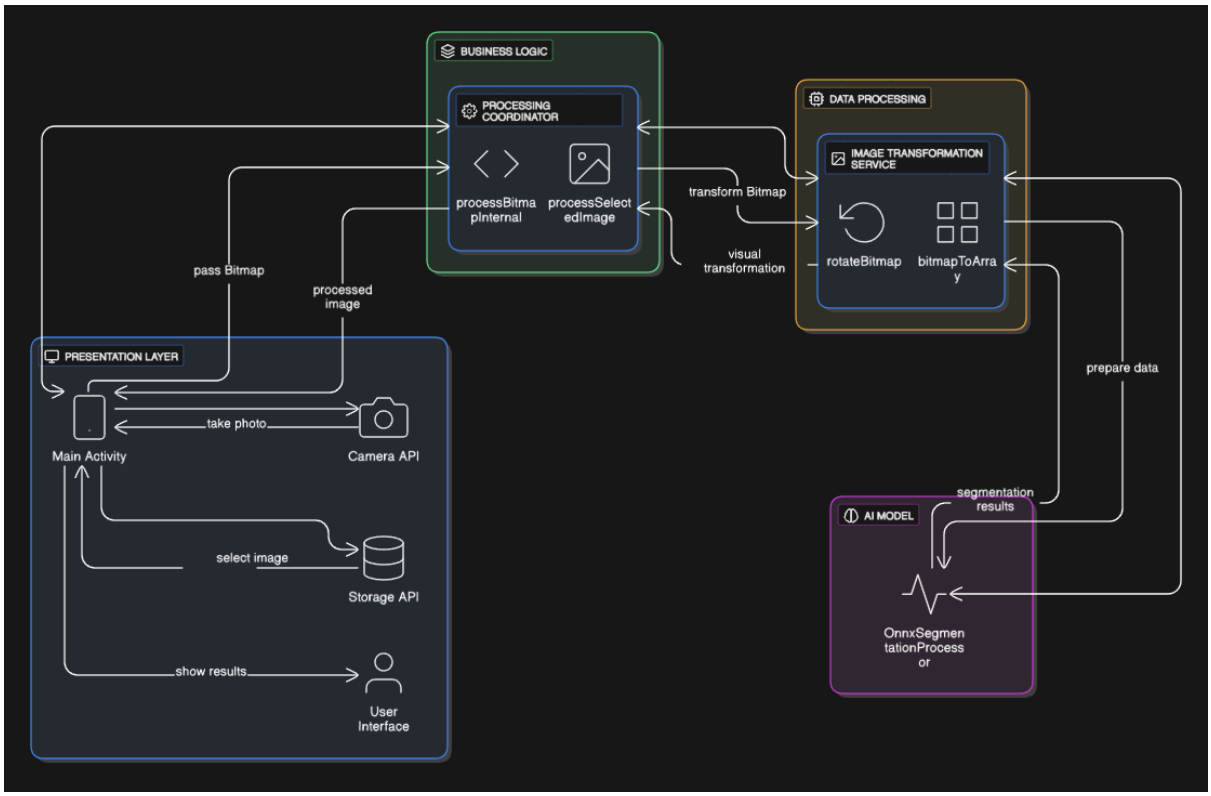

Figure 2: Diagram of the mobile application architecture

The interaction between the system components is implemented according to the principle of a unidirectional data flow. An image is first captured at the presentation layer and subsequently passed through the business logic layer to the data processing layer, where it is transformed into the appropriate format. The processed image is then analyzed at the machine learning model layer, after which the results are propagated back through the business logic layer to the presentation layer for visualization to the user. Data exchange between layers is facilitated through intermediary objects, such as Bitmap for image representation and Array for numerical data structures.

Owing to this architectural design, the system exhibits a high degree of modularity, allowing individual components to be developed and enhanced independently without necessitating substantial modifications to other parts of the system. Furthermore, the system ensures high performance even on devices with limited computational resources through the utilization of standard Android platform components and optimized libraries for image processing and machine learning.

4.2. Image Segmentation Model Training

4.2.1. Dataset preprocessing

The initial stage of the system development involved the creation and preprocessing of a specialized dataset. For this purpose, the ADE20K dataset was selected [10]. This dataset comprises a large collection of images annotated with 150 object classes, where each class is associated with a specific color in the image mask. For instance, class №1 (wall) corresponds to all pixels with the RGB value (1, 1, 1), class №2 (building) corresponds to pixels with RGB (2, 2, 2), and so forth.

A decision was made to reduce the number of classes from 150 to 20, focusing on objects most commonly encountered in indoor environments. To achieve this, a data structure in the form of a dictionary was created, grouping the original classes into 20 consolidated categories. Subsequently, an algorithm was developed to transform input images, which receives the image in its numerical representation.

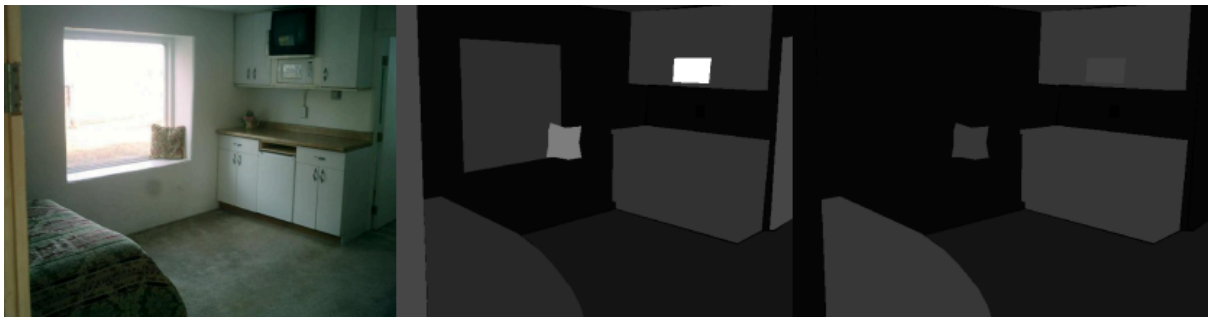

Figure 3: Image and mask before and after relabel

The numerical representation of an image is an array of numbers (three-dimensional for RGB) that defines the image resolution and color depth. For instance, a 640×540 RGB image produces an array of shape (640, 540, 3), where the last dimension indicates that each pixel contains three values corresponding to the red, green, and blue channels. These values range from 0 to 255, determining the intensity of each color channel and enabling the representation of any color.

The algorithm for reclassifying image masks uses the dictionary to map all original class RGB values to the corresponding new class values. Due to the reduction in the number of classes and their associated values approaching zero, the overall color palette visually darkened. For efficient storage and handling of the new image dataset, Apache Parquet format was selected, which maintains two columns containing the original images and their corresponding new masks.

4.2.2. Training of the Image Segmentation Model

After the creation and preparation of the dataset, the subsequent phase in the development of the system involved training the image segmentation model. Several primary architectures are commonly employed for image processing, among which U-Net [11], SegNet [12], and DeepLab [13] are most frequently utilized.

For this task, the SegFormer [14] architecture, was selected due to its demonstrated high efficiency in semantic segmentation tasks and its suitability as the foundation for our NVIDIA/MIT-B3 model. To facilitate model training, the dataset was partitioned into training, validation, and test subsets in a 70:15:15 ratio. This allocation ensures an optimal balance between effective learning and the ability to objectively evaluate the model's performance.

Since the task of semantic segmentation requires per-pixel classification of images, specific data transformation functions were implemented. For the training set, data augmentation techniques were applied, including random horizontal flips, adjustments of brightness, contrast, and saturation, as well as minor geometric distortions. These techniques enhance the model's robustness to variations in input data and help prevent overfitting.

To optimize the model parameters, the AdamW [15] algorithm was employed, which is a modification of the classical Adam optimizer with improved weight regularization. Training was conducted with an initial learning rate of 1e-4 and a scheduler that reduced the learning rate as the loss function stabilized. Analysis of the training and validation processes demonstrates a consistent improvement in performance metrics throughout the entire training period. The presented graphs visually illustrate the dynamics of key metrics and training parameters.

The training loss curve (train/loss) exhibits a rapid initial decrease from approximately 3.0 to 1.5 during the first iterations, indicating a swift adaptation of the model to the data. Subsequently, a gradual reduction in loss is observed, reaching values around 0.8, which reflects a consistent improvement in segmentation quality. The final loss value stabilized at approximately 0.66. The learning rate (train/learning_rate) was initialized at 0.0005 and gradually decreased to 0.0002 during the final stages of training. This strategy enabled the model to efficiently converge to optimal parameters while avoiding oscillations throughout the training process.

Figure 4: Train/loss chart

The validation results demonstrate a consistent improvement in the model's performance. The loss function on the validation set (eval/loss) decreased from initial values of approximately 1.1 to a final value of 0.7881, indicating good model generalization. Analysis of the curves shows that the model training proceeded steadily, with gradual improvements across all key metrics. The model achieved a balance between minimizing errors on the training set and maintaining generalization capability, as evidenced by the decreasing loss values during both training and validation phases. The final model exhibits consistently low loss values (train/loss = 0.66, eval/loss = 0.7881), reflecting a high quality of semantic segmentation.

4.3. System Functionality Description

The system's functionality is structured around three primary use-case scenarios, tailored to address the needs of the target user group. It enables the capture of environmental images through two main methods. First, the user can directly take a photograph using the device camera by selecting the "Capture Image" button within the application interface. During this process, the system automatically requests the necessary camera permissions if they have not yet been granted. Alternatively, the user may utilize the "Select Image" button to choose a previously stored image from the device gallery. This functionality is particularly advantageous for pre-examining specific locations prior to navigating them in real time.

Once an image is captured, the system automatically processes it to identify safe walkable surfaces. The system classifies various elements within the image, such as floors and other navigable areas, using a neural network. Users with partial visual impairment can quickly discern safe zones for movement by identifying the clearly marked semi-transparent green areas on the processed image. Following image analysis, the system provides textual guidance indicating how to safely reach the farthest point within the scene.

The system enables users to compare the highlighted safe zones with the actual surrounding environment by simultaneously displaying both the original and processed images on the device screen. This feature is particularly valuable for users with partial visual impairment, as well as for assisting personnel who may help fully blind users interpret the results effectively.

During image analysis, the system displays a progress indicator to inform the user of the ongoing processing status. In addition to the generated mask, detailed information derived from the analysis is presented to allow for more precise adjustment and understanding of the application's performance. Displayed parameters include the resolution of both the original and processed images, the shape of the input array, and the time required for processing, which can be used for statistical evaluation. This information enables users to assess the accuracy and reliability of the analysis results.

A key feature of the application is its capability to operate autonomously, analyzing images without requiring an internet connection. This ensures user data confidentiality and enhances usability in

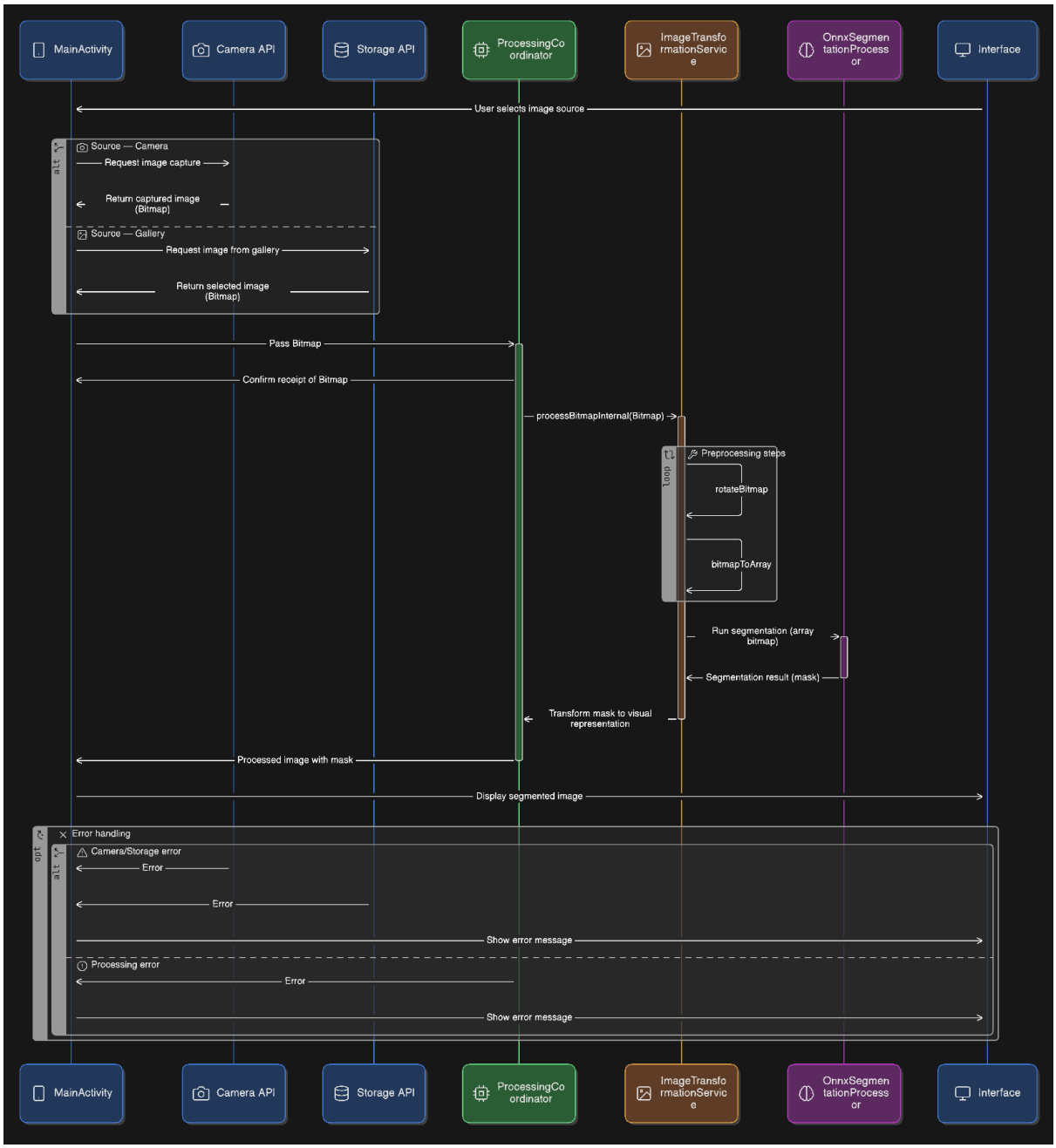

Figure 5: Sequence diagram of application

various environments, which is particularly critical for individuals with visual impairments who may navigate diverse and unpredictable conditions.

The system provides clear warnings and actionable recommendations to address issues that may arise during its use. For instance, if camera access permission is denied, the system notifies the user that granting permission is required to proceed. Similarly, if image processing fails, the system alerts the user to the issue and suggests attempting an alternative image. This proactive feedback mechanism enhances usability and ensures that users can effectively navigate potential obstacles while interacting with the application.

Figure 6: An example of how the segmentation model works

5. Demonstration of Concept Feasibility

Figure 6 illustrates the system's performance in typical indoor environments. As shown in the presented images, the system successfully segments the floor (highlighted in bright green) across various scenarios. The upper pair of images demonstrates segmentation results where the floor is clearly distinguished from other indoor objects, including furniture and walls. The lower images depict the original scenes without the overlay of the segmentation mask.

When deployed on the mobile device, the model exhibits sharper edges at the junctions between different object classes, which is attributed to weight compression during the model transfer process. The figure also shows a comparison of the original image and the segmentation mask highlighting the safe surface for navigation, alongside directional instructions generated by the system, such as: "Move 0.5 meters to the right and walk straight 4.17 meters." This instruction is the output of the find_optimal_path algorithm. During testing on a dataset of 50 images from various indoor environments, the system achieved an Intersection over Union (IoU) metric of approximately 70%, indicating a robust segmentation accuracy.

6. Conclusions

As a result of this work, a mobile application has been developed that leverages computer vision technologies to assist visually impaired individuals in safely navigating indoor environments without the use of a white cane.

A key achievement of this study is the development of a surface recognition model for safe navigation, achieving an approximate accuracy of 70%. In addition to the segmentation model, an algorithm for distance estimation within images was implemented, allowing users to approximate the distance to the farthest point along the intended path of movement.

Testing the application in real indoor environments demonstrated high accuracy in identifying safe pathways: over 90% in scenarios without small obstacles and over 70% when obstacles were present.

The system operates autonomously, processing images directly on the mobile device, thereby ensuring data privacy and convenient use in a variety of conditions.

However, the current implementation exhibits several limitations:

1. The application does not support voice commands, which significantly complicates user interaction.
2. The system lacks the functionality to provide auditory instructions for safe navigation trajectories, further limiting ease of use.
3. Users are required to manually capture images of the environment with the camera at intervals; a more practical solution would involve real-time image processing to enable continuous navigation.

Addressing these limitations constitutes a primary direction for future research in this domain.

It is also important to outline the following steps aimed at improving the performance of the application and enhancing the overall quality of the research:

1. Limited test dataset - the current test set is relatively small and does not include class-based metrics or confidence/uncertainty estimations. Future research will focus on expanding the test dataset and implementing classification of partially visible objects
2. Distance estimation improvement - it is necessary to investigate more accurate methods for determining distances between objects, as the current approach, which relies on the standard human height and device tilt angle, remains rather approximate
3. Field testing and user evaluation - after the development of the MVP version of the application, it is planned to publish it for real-world testing with blind and visually impaired participants to assess usability, safety, and task performance

.

It is also important to note that, as the application has been developed solely at the proof-of-concept stage, it has not yet been tested with visually impaired individuals. Conducting such user trials represents a critical priority for subsequent studies to assess both usability and effectiveness.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to Grammar and spelling check. After using these tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] Blindness and vision impairment, World Health Organization (2023). URL: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

[2] Z. R. Kahaki, M. Karimi, M. Taherian, R. Simi, Development and validation of a white cane use perceived advantages and disadvantages (wcpad) questionnaire, BMC Psychol. 11(1) (2023). doi:10.1186/s40359-023-01282-4.

[3] W. Agnew, M. Cheng, Computer-vision research powers surveillance technology., Nature (2025) 73–79. doi:10.1038/s41586-025-08972-6.

[4] Screen Reader User Survey 10 Results, 2024.

[5] Helping blind and low vision people, Be My Eyes, 2025. Available: https://www.bemyeyes.com/.

[6] Seeing AI, Microsoft, 2025. Available: https://www.microsoft.com/en-us/ai/seeing-ai.

[7] P. Jiang, D. Ergu, F. Liu, Y. Cai, B. Ma, A review of yolo algorithm developments, Procedia Computer Science 199 (2022) 1066–1073. URL: https://www.sciencedirect.com/science/article/pii/S1877050922001363. doi:https://doi.org/10.1016/j.procs.2022.01.135, the 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 2021): Developing Global Digital Economy after COVID-19.

[8] P. Bharati, A. Pramanik, Deep learning techniques—r-cnn to mask r-cnn: A survey, in: A. K. Das, J. Nayak, B. Naik, S. K. Pati, D. Pelusi (Eds.), Computational Intelligence in Pattern Recognition, Springer Singapore, Singapore, 2020, pp. 657–668.

[9] S. Hao, Y. Zhou, Y. Guo, A brief survey on semantic segmentation with deep learning, Neurocomputing 406 (2020) 302–321. URL: https://www.sciencedirect.com/science/article/pii/S0925231220305476. doi:https://doi.org/10.1016/j.neucom.2019.11.118.

[10] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ade20k dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[11] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[12] S. Alqazzaz, X. Sun, X. Yang, L. Nokes, Automated brain tumor segmentation on multi-modal mr image using segnet, Computational Visual Media 5 (2019) 209–219. doi:10.1007/s41095-019-0139-y.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (2018) 834–848. doi:10.1109/TPAMI.2017.2699184.

[14] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, Segformer: Simple and efficient design for semantic segmentation with transformers, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 12077–12090. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

[15] P. Zhou, X. Xie, Z. Lin, S. Yan, Towards understanding convergence and generalization of adamw, IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (2024) 6486–6493. doi:10.1109/TPAMI.2024.3382294.