FMDB Transactions on Sustainable Computing Systems

Innovative Technology for Enhancing Independence of the Blind or Visually Impaired

Abeer A. Amer[1, *], Esraa E. Mostafa[2], Dina N. Ahmed[3], Nourhan A. Mahmoud[4]

[1]Department of Computer Science and Information System, Faculty of Management Sciences, Sadat Academy for Management and Sciences, Alexandria, Egypt.
[2,3,4]Department of Computer and Data Science, Alexandria University, Alexandria, Egypt.
abamer.2000@gmail.com[1], cds.esraaelsayed75187@alexu.edu.eg[2], cds.dinanabil14073@alexu.edu.eg[3], cds.nourhanabdallah12988@alexu.edu.eg4

Abstract: People with visual impairments face challenges in tasks like object recognition and text reading. SmartSight, leveraging Computer Vision technology, aims to replicate human visual perception to overcome these hurdles. The device seamlessly combines a compact camera and a dedicated computing unit with a mobile application, offering indispensable assistance to the visually impaired. This innovative solution excels in object detection, facial recognition, text-to-speech capabilities, food identification, environmental description, and even currency recognition, specifically Egyptian currency. SmartSight's integration with a mobile application enhances user control and customization. Its compact design emphasizes portability and ease of use. By providing a comprehensive set of functionalities, SmartSight significantly improves the daily lives of individuals with visual impairments, offering them newfound independence and accessibility. The device's ability to identify objects, read text aloud, recognize faces, describe surroundings, and discern currency demonstrates a holistic approach to addressing the diverse needs of the visually impaired. In summary, SmartSight stands out as a transformative solution, symbolizing a substantial leap in enhancing the quality of life for those living with visual impairments.

Keywords: Blind Assistance System; Computer Vision; Visual Impairments; Mobile Application; Intricate Algorithms; Mathematical Models; SmartSight's Integration; Independence and Accessibility; Facial Recognition.

Received on: 27/04/2023, **Revised on:** 19/08/2023, **Accepted on:** 11/10/2023, **Published on:** 20/12/2023

Cite as: A. A. Amer, E. E. Mostafa, D. N. Ahmed, and N. A. Mahmoud, "Innovative Technology for Enhancing Independence of the Blind or Visually Impaired," *FMDB Transactions on Sustainable Computing Systems*., vol. 1, no. 4, pp. 204–212, 2023.

1. Introduction

On a global scale, the World Health Organization reports an estimated 285 million individuals grappling with visual impairment. Among them, 246 million contend with varying degrees of visual impairment, while 39 million grapple with complete blindness. Notably, the Eastern Mediterranean region contributes significantly to the global prevalence of blindness, accounting for approximately 12.6% of cases. In the specific context of Egypt, an estimated 3.5 million individuals have blindness, with roughly 37,000 blind individuals actively engaged in various educational pursuits [1]. Visual impairment and blindness present individuals with daily obstacles. Computer vision technology holds the potential to revolutionize their experiences, offering increased independence, safety, and access to information. This field of artificial intelligence equips machines with the ability to interpret visual data through intricate algorithms and mathematical models, finding applications in areas such as object and facial recognition, autonomous vehicles, and medical imaging, among others. Recent advancements in technology underscore the growing significance of computer vision, with the promise of enhancing various facets of our daily lives.

[*] Corresponding author.

The present document introduces a comprehensive system consisting of a compact camera, a dedicated computing unit, and a mobile application, all meticulously designed to empower and improve the lives of individuals facing visual impairments. The goal is to reduce their reliance on external assistance. The document is structured as follows: Section 2 provides an overview of existing applications available in app stores. Section 3 outlines the methodology, while Section 4 delves into the fundamental algorithm and its practical implementation. Section 5 briefly summarizes the results obtained and potential avenues for improvement, with Section 6 serving as the conclusion of the paper.

2. Literature Review

Numerous mobile applications have been developed across various platforms, including Android, iOS, and Windows, with the primary objective of aiding visually impaired and blind individuals. These applications can be categorized into distinct groups, each catering to specific needs. These categories encompass object detection, text-to-speech functionality, currency recognition, facial recognition, and environmental scene description.

2.1. Related Mobile Applications

SuperSense: Available on Android and iOS, SuperSense employs artificial intelligence for tasks like text, currency, and product recognition. It provides camera guidance and features a smart scanner for automatic detection [2].

Lookout: Compatible with Android and iOS, Lookout utilizes AI for text-to-speech, currency identification, and product recognition. It offers camera guidance and an automated smart scanner [3].

2.2. Related Products

OrCam MyEye: A wearable assistive device attached to eyeglasses or sunglasses. It uses AI for reading text, facial recognition, and object identification. The built-in camera captures surroundings, with real-time analysis providing audio feedback through a speaker or headphones [4].

Aira: A technology company integrating wearable tech, AI, and human agents. The flagship product consists of smart glasses with a camera, enabling users to connect with Aira agents for real-time verbal guidance and assistance [5].

Envision An AI-based product enhancing daily life for visually impaired individuals. It utilizes computer vision and machine learning for real-time recognition and description of objects, text, and people. Envision excels in conveying text content, identifying colors, currencies, and emotions, and includes a voice assistant for user guidance and feedback [6].

Comprehensive Comparative Analysis: SmartSight vs Competitors

2.3. Facial Recognition

- SmartSight: Excels in advanced facial recognition, aiding social interactions.
- Competitors: OrCam MyEye and Envision include facial recognition, providing a more holistic user experience.

SuperSense, Lookout, and Aria do not offer this capability.

2.4. Text Reader

- SmartSight: Converts written content to audio, facilitating seamless information access.
- Competitors: All competitors share this feature, ensuring basic text accessibility.

2.5. Object Detection

- SmartSight: Excels in real-time object detection, providing enhanced environmental awareness.
- Competitors: Lookout, OrCam MyEye, and Envision offer object detection capabilities. SuperSense and Aria lack this feature, limiting their scope.

2.6. Environmental Description

- SmartSight: Provides detailed scene descriptions, enhancing user situational awareness.

- Competitors: Lookout and Envision offer scene descriptions, while SuperSense, OrCam MyEye, and Aria do not cover this aspect.

2.7. Money Recognition

- SmartSight: Pioneers in recognizing currency notes and fostering financial independence.
- Competitors: OrCam MyEye and Envision include cash recognition. SuperSense, Lookout, and Aria lack this vital feature.

2.8. General Food Recognition

- SmartSight: Innovates with food recognition, ensuring a broader understanding of the user's surroundings.
- Competitors: None of the listed competitors incorporate food recognition, highlighting SmartSight's uniqueness in addressing diverse user needs.

In summary, SmartSight emerges as a standout solution, offering a comprehensive feature set that encompasses object detection, text and scene description, cash and face recognition, and even food recognition. While some competitors excel in specific functionalities, SmartSight provides a more inclusive and versatile assistive technology solution for individuals with visual impairments.

3. Methodological Approach

The development of a blind assistance system involves an organized methodology that encompasses several key stages, starting with the definition of system requirements, followed by architectural design, algorithm selection, implementation, testing, evaluation, and iterative refinement. Our approach to this process involves a comprehensive division into software, hardware, mobile applications and their subsequent integration. Each component undergoes a sequence of model building, rigorous testing, and data collection to enhance accuracy continually. In this setup, the mobile application serves as the control device, while the hardware includes a computer and camera for executing various features. The subsequent sections within the document delve into the detailed features, algorithms, and techniques applied in the implementation process, as depicted in Figure 1.

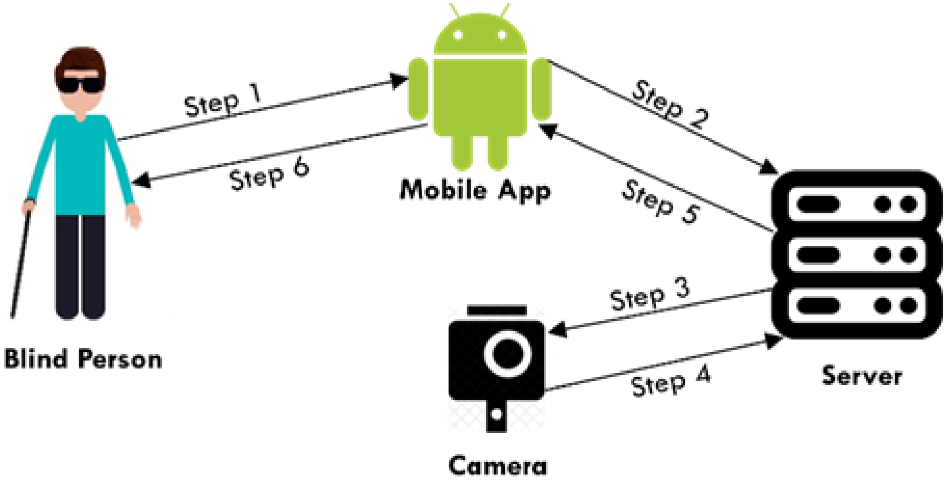

Figure 1: System Methodology

3.1. Feature 1: Facial Recognition

Face recognition is the technology used to identify individuals by analyzing their facial features in photos, videos, or real-time scenarios. It plays a crucial role in public safety, authentication, and retail. The process consists of two main steps: face detection and recognition. Face detection can be feature-based (identifying specific facial features) or appearance-based (recognizing facial patterns). Modern approaches often combine both methods for accuracy and speed. Subsequently, recognized faces are compared to a dataset of stored images to determine a person's identity, involving the encoding of images into pixels and pattern analysis, as shown in Figure 2 [7].

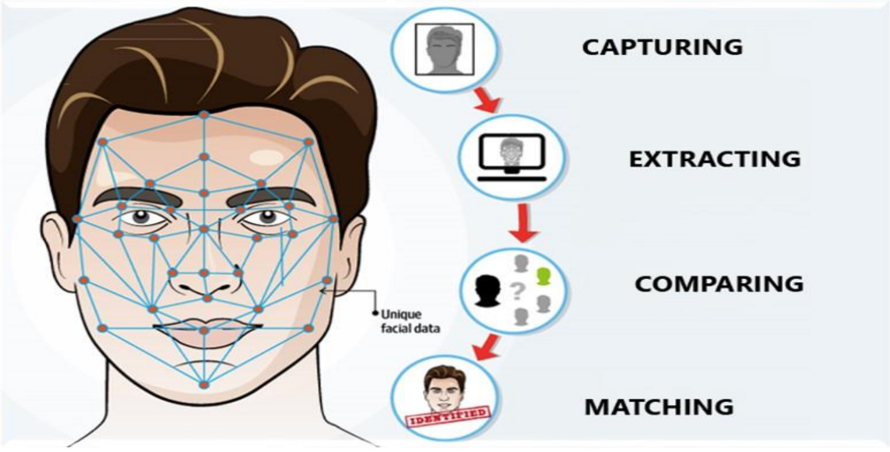

Figure 2: Facial Recognition

3.2. Feature 2: Text Reader

Text recognition holds paramount significance in the process of transforming text images into machine-readable characters or complete sentences. This intricate process encompasses two key aspects:

Character Recognition: Character recognition involves the segmentation of text images into individual characters, typically achieved through Optical Character Recognition (OCR) techniques. Word Recognition: Word recognition takes the output from character recognition and combines it with language models to interpret and generate complete words. Tesseract OCR, an open-source engine, stands as a versatile and adaptable solution for text recognition. Notably, it supports multiple languages and provides the capability to extract text from images via Application Programming Interfaces (APIs). Tesseract OCR can be employed both independently and in conjunction with external text detection mechanisms, offering flexibility in recognizing text within diverse contexts. This versatility positions it as a valuable tool with wide-ranging applications, as illustrated in Figure 3 [8].

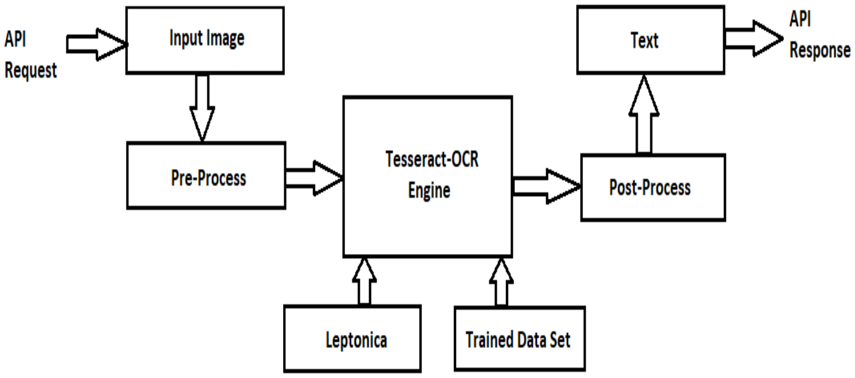

Figure 3: OCR Process Flow

Tesseract 4.00 introduces a novel neural network subsystem designed for text line recognition. This subsystem was originally inspired by OCRopus's Python-based Long Short-Term Memory (LSTM) implementation but has been entirely reengineered in C++. Notably, this subsystem is compatible with TensorFlow through Vectorized Graphic Shading Language (VGSL). In the context of character recognition, Convolutional Neural Networks (CNN) are commonly employed, while text sequences are effectively managed using Recurrent Neural Networks (RNN), with a particular emphasis on Long Short-Term Memory (LSTM) networks. Tesseract's OCR model incorporates LSTM networks for sequence learning; however, it's worth noting that the computational load can increase with a high number of states. Tesseract's foundational roots trace back to OCRopus, a

Python-based model that itself evolved from CLSTM, which is a C++ implementation of LSTM. CLSTM relies on the Eigen library for numerical computations, as depicted in Figure 4 [9].

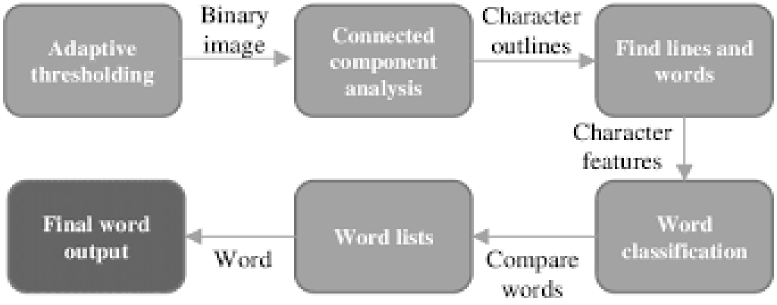

Figure 4: How the Tesseract OCR Model Works

3.3. Feature 3: Object Detection

Object detection, a prominent computer vision technique, serves to identify and precisely locate objects within images or videos. This capability supports various tasks, including object counting, accurate tracking, and object labeling. The process involves combining both object classification and localization, enabling the simultaneous recognition of multiple objects, each delineated by bounding boxes within the image, as illustrated in Figure 5 [10].

Figure 5: Computer Vision Tasks

The SSD (Single Shot Detector) model represents an efficient methodology for object detection, featuring three essential components:

Single Shot: This approach achieves object localization and classification within a single network pass, streamlining the overall process.

Multi-Box: It employs bounding box regression techniques to ensure precise and accurate object localization.

Detectors: These convolutional filters play a pivotal role in generating object category scores within default boxes. They adaptively adjust these boxes to enhance object fitting, and multiple feature maps are utilized to accommodate objects of varying sizes. Notably, SSD eliminates the need for proposal generation and resampling stages, rendering it particularly suitable for systems requiring object detection, as illustrated in Figure 6 [11].

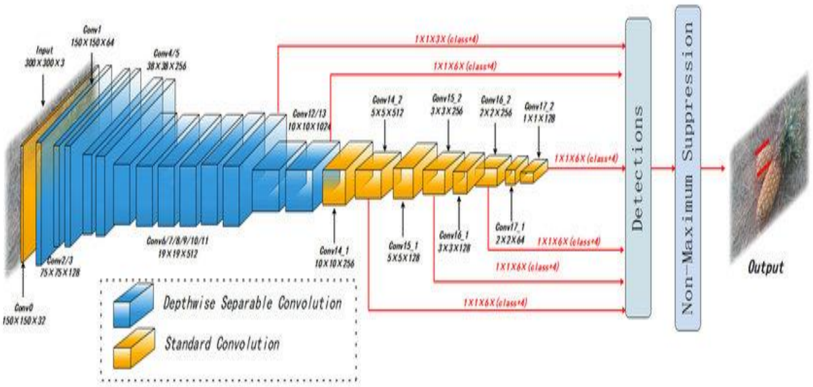

Figure 6: MobileNet & SSD

3.4. Feature 4: Environmental Description (Image to text)

Image description, or 'image captioning,' combines Natural Language Processing and Computer Vision to create textual descriptions for images, benefiting the visually impaired. This method integrates Convolutional Neural Networks (CNN) for feature extraction, often using models like Inception trained on ImageNet. These features are then processed by Long Short-Term Memory (LSTM) networks to generate image captions. The CNN functions as an encoder, and the last hidden state connects to the decoder for caption generation, enhancing the understanding of image content for people who are blind, as shown in Figure 7 [12].

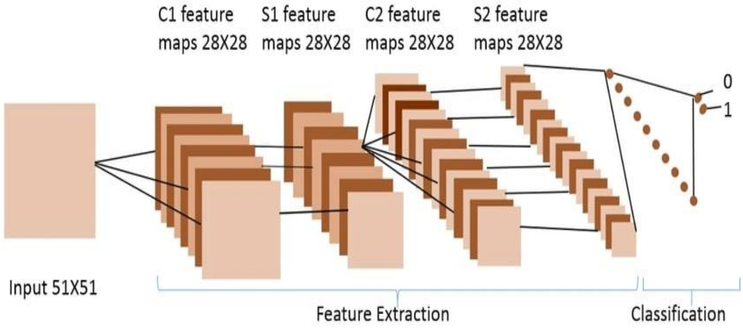

Figure 7: Feature Extraction

The Flickr30k dataset consists of 31,783 images, each associated with five descriptive captions, making it a valuable benchmark for sentence-based image descriptions. This dataset finds widespread application in the realm of associating linguistic expressions with visual content. The Decoder, implemented as a Recurrent Neural Network (RNN), is responsible for word-level language modeling and cooperates with the Encoder. LSTM, a specialized RNN variant, excels in sequence prediction tasks, rendering it apt for functions like next-word predictions and information management during input processing, as depicted in Figure 8 [13].

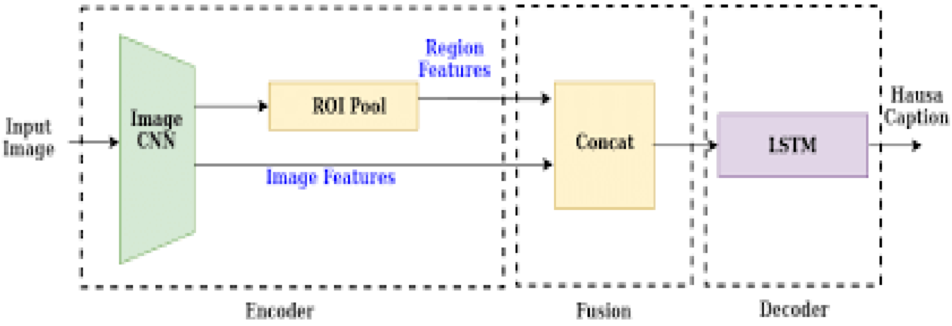

Figure 8: Caption Generator

3.5. Feature 5: Money and General Food Recognition (Image Classification)

Image classification is a task within the field of computer vision that involves the assignment of labels to images. These labels can span a range from broader categories like 'cat' or 'dog' to more specific objects such as 'Siamese cat' or 'Golden Retriever.' This task is particularly challenging due to the inherent complexity of images that may contain various objects. In our specific context, we utilize image classification to categorize food and currency, making use of a Convolutional Neural Network (CNN) architecture. In this framework, each category of currency serves as a distinct class. Image classification is fundamentally a supervised learning task, where pre-labeled training data is employed to train a machine learning algorithm.

This training process imparts the ability to recognize visual features and subsequently classify unlabeled images based on these learned features. The process inherently involves feature extraction to identify and capture meaningful patterns within the dataset. Transfer learning is a machine learning technique wherein a model initially developed for one specific task serves as the foundational basis for creating a model intended for a different task. This method is frequently employed in deep learning. Pre-trained models, crafted by experts and trained on large and diverse datasets to tackle analogous problems, are pivotal components of this approach. By utilizing pre-trained models, AI teams can initiate their work with a well-established model rather than constructing one from the ground up, as depicted in Figure 9 [14].

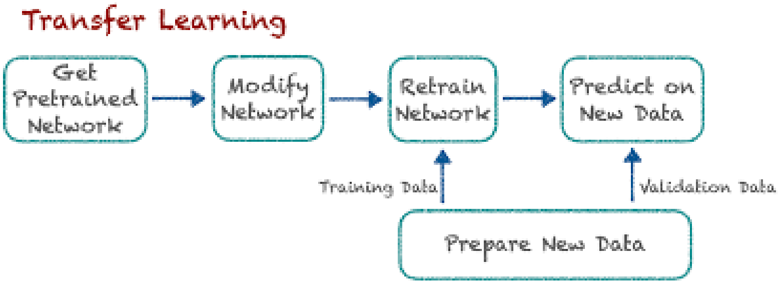

Figure 9: Transfer Learning

4. Proposed System and Implementation

4.1. Proposed System

The system comprises two primary components: a mobile phone, which acts as a control device through an application, and a mini-computer, responsible for executing system functions. The user initiates interaction by selecting a feature on the mobile app, signified by a confirmation sound. The request is transmitted to the mini-computer, which, in turn, directs camera-equipped glasses to capture an image. Following image processing and text-based output generation, the results are relayed back to the app for conversion into sound, ultimately presented to the user, as depicted in Figure 10.

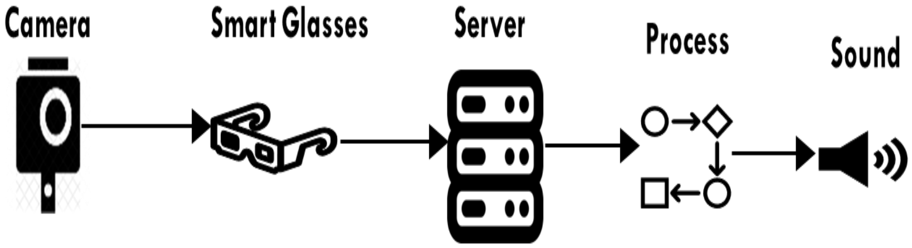

Figure 10: Proposed System

4.2. Results and Discussion

In this study, we developed a system that leverages key tools, libraries, and hardware components to perform computer vision tasks. The system utilizes PyTorch, TensorFlow, and OpenCV for training computer vision models. These libraries provide a wide range of functionalities for image processing, feature extraction, and model training.

To enhance the accuracy of object detection and text reading, we integrated the Google Cloud API into our system. This API provides advanced algorithms for object detection and optical character recognition (OCR), which significantly improve the performance of our computer vision models.

The processing of images takes place on a Raspberry Pi mini-computer. This compact device is equipped with a mini camera for image capture. The Raspberry Pi offers sufficient computational power to perform real-time image processing tasks while being energy-efficient.

User control is facilitated by a mobile app developed using Flutter. This cross-platform framework allows us to create an intuitive user interface that can be easily accessed on both Android and iOS devices. The app provides users with the ability to select desired features executed on the Raspberry Pi.

To provide audio output, we integrated TTS (Text-to-Speech) functionality into our mobile app. This allows the system to convert text-based information into spoken words, providing an additional layer of accessibility for visually impaired users or situations where visual feedback is not possible.

Furthermore, we incorporated the Google Translate API into our system to enable multilingual support. This API allows us to translate text from one language to another in real time, expanding the usability of our system across different language barriers. Communication between the mobile app and Raspberry Pi is enabled through the Python Socket library. This library provides a convenient way to establish a network connection between two devices and exchange data seamlessly. It ensures smooth communication between the user interface on the mobile app and the computational capabilities of the Raspberry Pi.

Overall, our system demonstrates effective integration of various tools, libraries, and hardware components to create a robust computer vision solution. By leveraging state-of-the-art machine learning frameworks like PyTorch and TensorFlow along with powerful APIs such as Google Cloud Vision API and Google Translate API, we were able to achieve high accuracy in object detection and text reading tasks.

The use of Raspberry Pi as a processing unit offers flexibility in terms of deployment scenarios due to its small form factor and low power consumption. Additionally, by developing a mobile app using the Flutter framework, we ensured compatibility across multiple platforms while providing an intuitive user interface for controlling system features.

The integration of TTS functionality enhances accessibility by providing audio output for visually impaired users or situations where visual feedback is not feasible. Furthermore, multilingual support through Google Translate API expands the usability of our system across different language barriers.

In conclusion, our system successfully combines cutting-edge tools, libraries, and hardware components to create an efficient computer vision solution with enhanced object detection accuracy and multilingual support. The integration of various technologies enables seamless communication between user control via mobile app and image processing on the Raspberry Pi mini-computer.

5. Conclusion

Our computer vision system emerges as a transformative and empowering tool for blind and visually impaired individuals, transcending the limitations imposed by their condition. The comprehensive set of features, ranging from object detection to text reading, scene description, face recognition, cash recognition, and food identification, underscores our commitment to enhancing their autonomy and daily experiences. The user-centric design of the system, seamlessly integrated with a user-friendly mobile app, facilitates intuitive control through voice commands. This not only simplifies the interaction but also ensures that individuals with visual impairments can independently harness the capabilities of the technology. The inclusion of unique features like food and cash recognition is particularly noteworthy, as it eliminates the reliance on external assistance, fostering a greater sense of self-reliance. The system's compact components, a mini-camera, mini-computer, and purpose-built mobile app work in harmony to provide a holistic solution. By addressing varied needs, from recognizing faces to discerning currency, the system is poised to significantly improve the overall quality of life for visually impaired individuals. It goes beyond mere functionality; it symbolizes a commitment to inclusivity, accessibility, and the fundamental right of every individual to live independently and participate fully in the world around them. In essence, our computer vision system stands as a testament to the potential of technology to break barriers, offering a brighter and more inclusive future for those with visual impairments.

Acknowledgment: The authors wish to thank the other members, Mazen M. El-Sayed and Mohamed E. El-Sayed.

Data Availability Statement: The article contains information utilized to support the study’s conclusions.

Funding Statement: No funding was used to write this manuscript and research paper.

Conflicts of Interest Statement: No conflicts of interest exist, according to the authors, with the publishing of this article.

Ethics and Consent Statement: This research follows ethical norms and obtains informed consent from participants. Confidentiality safeguards protected privacy.

References

1. B. Bourne, S. Ackland, and H. Resnikoff, "World blindness and visual impairment: A review," Eye, vol. 29, no. 1, pp. 79-89, 2015.
2. SuperSense, "Open Sight," [Online]. Available: https://www.opensight.org.uk/supersense-app-review. [Accessed: 03-Mar-2023]
3. L.-A. Vision, "Google Play," [Online]. Available: https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en_US. [Accessed: 03-Mar-2023]
4. O. MyEye, "OrCam," [Online]. Available: https://www.orcam.com/en-us/orcam-myeye. [Accessed: 03-Mar-2023]
5. Aira, "Aira," “We’re Aira, a visual interpreting service,” Aira - Visual Information On Demand, 29-Sep-2021. [Online]. Available: https://aira.io/. [Accessed: 03-Mar-2023].
6. Envision, "Letsenvision," [Online]. Available: https://www.letsenvision.com/glasses. [Accessed: 03-Feb-2023]
7. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, 2015.
8. L. Wang and X. Yang, "CRNN: Convolutional Recurrent Neural Network for Text Recognition," Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 1381-1389, 2017.
9. L. Wang, W. Li, W. Liu, and J. Zhang, "CRAFT: Character Region Awareness for Text Detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA, pp. 254-263, 2018.
10. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Object Detection in Computer Vision: Techniques and Applications," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, pp. 1801-1813, 2015.
11. S. Zhang, G. Lin, and J. Zhang, "Single Shot MultiBox Detector," Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, pp. 21-37, 2016.
12. A. Karpathy, L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 3128-3137, 2015.
13. F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," Neural Computation, vol. 12, no. 10, pp. 2451-2471, 2000.
14. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, Lake Tahoe, NV, pp. 1097-1105, 2012.