Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

YOLOv12-Based Multi-Feature Detection Framework with Voice Assistant for Enhanced Mobility and Independence of Visually Impaired Persons

FIRST A. AUTHOR[1], (Fellow, IEEE), SECOND B. AUTHOR[2], and Third C. Author, Jr.[3], (Member, IEEE)

[1]National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@boulder.nist.gov)
[2]Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu)
[3]Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA

Corresponding author: First A. Author (e-mail: author@ boulder.nist.gov).

This paragraph of the first footnote will contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

ABSTRACT

The provision of safe, independent and context-aware mobility for people with visual impairment is still an open issue, especially in resource-constrained or region-specific scenarios. In this paper, we propose a deployment-ready and trainable together with real-time voice assistant YOLOv12-based multi-feature assistive vision framework to support navigation, environment awareness and user autonomy. It is distinctive from existing assistive systems that primarily targets standalone perception tasks; the proposed modular, single-lightweight pipeline integrates environment object detection, footpath safety classification, currency recognition in Bangladesh currency, face identification (both known and unknown), and bilingual (Bangla-English) optical character recognition. The framework exploits the attention-based YOLOv12 for strong trade off between detection accuracy and real-timing performance, which is particularly apt for CPU based edge and wearable devices. In this work, we aim to alleviate the shortcomings of publicly available generic datasets, and propose several task-driven region-specific datasets that contain culturally, linguistically and environmentally relevant data. This is validated through comprehensive experimental results achieving reliable and consistent performance across all assistive tasks: a currency recognition mAP@0.5 0.951 and reliable footpath safety evaluation, robust objects detection and high face recognition rate. An integrated voice assistant transfers multi-modal perception outputs to natural spoken instructions and gives rise to hands-free and low-cognitive-load interactions. In summary, the proposed system represents a scalable, practical and socially meaningful assistive solution to increase mobility, safely and independence for visually impaired individuals in everyday environments.

INDEX TERMS YOLOv12, Object Detection, Footpath Detection, Banglesh Currency Detection, Known and Unkown Face Recognition, OCR.

I. INTRODUCTION

VISUAL impairment is still a significant public health and social problem worldwide, limiting independent travel as well as daily life. The World Health Organization (WHO) reports that more than 2.2 billion people globally suffer from visual impairment, a substantial portion of whom encounter significant daily challenges in walking through environments safely, avoiding obstacles, understanding what they are looking at and accessing printed information [1]. The lack of visual cues can cause potential severe personal safety and the loss of autonomy or dignity in unknown or dynamic environments. The blind or visually impaired typically use mobility aids such as white canes or guide dogs. Cost-effective solutions like LIDAR can help in obstacle detection and spatial localization, but they may not be semantic enough (like specifying the identity of objects and their spatial relations or some contextual information) [2]. As a result, such individuals frequently rely on the help of others

(e.g., reading signs, money recognition or face recognition), thereby affecting their independence and quality of life [3].

Assistive devices have developed from simple mechanical support to so called electronic travel aids (ETAs) using ultrasonic sensors, infrared and wearable devices in combination with haptic or auditory feedback in recent years [4]. While some of these systems are able to perceive obstacles in their immediate vicinity, most do not have the ability to understand complex visual scene or give high-level semantics. This inability has inspired increasing attention to the vision-based assistive technologies, which utilize camera sensors and computer vision algorithms for filling perception gaps for people with visual impairment.

Deep learning and computer vision have made substantial progress in enhancing the ability of computers to perceive visual world. Convolutional neural networks (CNNs) achieved great success in image classification, object detection and scene understanding which can generative reliable visual perception even in the complicated real-world settings [5]. Such advances have influenced vision-based assistive systems by facilitating richer situational understanding and more robust perception-driven assistance.

In modern single shot detectors, You Only Look Once (YOLO) becomes popular for its unified and real-time detection and computational efficiency [6]. YOLO formulates object detection as a regression problem for the bounding box, and performs both localisation and classification at the same time in one forward pass. There have been new versions, which are YOLOv4 and YOLOv7 since then that were making architectural optimizations resulting in higher detection accuracy yet real- time processing performance retained, suitable for assistive navigation applications [7], [8]. Far beyond object detection, other assistive tasks are also ameliorated by the approaches of deep learning including OCR, face recognition and scene parsing. Numerous recent assistive systems combine sensing visual perception with auditory or haptic feedback in order to provide visually impaired users with meaningful information about their environment [9], [10]. However, existing methods are mostly limited to individual functions and do not provide an overall perception framework that can deal with several assistant tasks at the same time.

However, despite such progresses, existing assistive navigation systems remain to be constrained in several aspects. Several systems have been developed for specific tasks like reading and obstacle avoidance in isolation, without taking into account the interdependency of perception during real-world navigation [11], [12]. Furthermore, many existing works are based on generic public datasets which do not cover local-specific content such as local currency, language-specific text, or culturally relevant environment context, resulting in poor generalization ability when applied into practice [18]. Another profound difficulty is to balance the detection accuracy and real-time performance effectively. Many of the state-of-arts deep learning models, however, are computationally heavy and not feasible in practice on resource-constrained wearable or mobile platforms associated with assistive technologies [13]. User-centred studies further highlight the necessity of integrated multi-feature perception combined with timely, intuitive feedback to enhance usability, safety, and user trust [14].

Real-time auditory interaction is also important for assistive navigation of blind subjects, as well as vision perception [18]. Using feedback in the form of voice supports hands-free and low-cognitive-load transmission of contextual information, guiding users without interrupting their mobility tasks. Previous studies have also demonstrated that auditory feedback greatly increases situational awareness and a natural human navigating characteristic, navigation efficiency in assistive systems [9], [10]. As a result, the coordination of voice assistant and vision modules becomes necessary in order to provide timely and easy-to-use help among dynamic environments.

Inspired by these challenges, the primary objective of this work is as follows: To develop a multi-feature detection framework based on YOLOv12 combined with a real-time voice assistant with the capability to improve environmental perception and navigation assistance for visually impaired users. Unlike the previous single-task works, this framework combines a large number of vision-based modules in an unified pipeline such as environmental object detection, Bangladesh currency recognition and footpath safety classification, identity verification with known and unknown face and Bangla–English OCR. Obstraction detection can help to realize obstacle avoidance and safe driving [9], [10], currency recognition can provide a chance for people living in remote areas making independent financial system access [15]. Footpath analytics and face familiarity recognition contribute to enhanced pedestrian safety, social awareness, while OCR supports instantaneous access to the textual information (e.g., signboards, warning signs and room labels) [16]. The built-in voice assistant turns these perception outputs into live auditory feedback, providing convenient hands-free interaction.

For good support of these tasks, a proprietary data set was developed, since the existing public data sets do not have sufficiently comprehensive coverage for region related objects, local currency value notes and text in language specific writing and realistic navigation situations relevant for the target user group. We focused on task relevance, annotation quality and environmental realism instead of dataset scale which is consistent with some latest work in the field of assistive vision [17]. The proposed framework is developed considering online operation, robustness and portability to embedded platforms such as wearables or mobiles. While the system contains region-dependent modules, through modularity it is portable to other languages, currencies and deployment scenarios providing an experience-based simplification tool for assistive wayfinding systems internationally [18].

CONTRIBUTIONS

Some of the main contributions of this work are:

- We present an all-in-one YOLOv12-based multi-feature assistive vision system with real-time voice assistant for

environment object detection, footway safety status assessment, Bangladesh currency recognition, known and unknown face identification and Bangla-English optical character recognition.

- We curate task-specific and region-specific datasets designed for assistive navigation, which supplement the deficiencies of existing generic public datasets in the dimension of cultural relevance, local currency representation, linguistic diversity and real-world environmental challenge.
- We develop a light-weight and attention- enhanced detection pipeline which strikes a good balance in accuracy and real-time behaviour, making the detection available for reliable deployment on CPU based edge and wearable platforms without the necessity of dedicated hardwareacceleration.
- We perform thorough quantitative and qualitative assessments over various assistive tasks, evidencing the robustness, practicality and real-world value of our proposed system in enhancing mobility, safety and situational awareness of visually impaired users.

II. LITERATURE REVIEW

Nowadays, with the rapid development of computer vision, deep learning and edge computing technologies, researches on assistive technology for visually impaired people have become more and more popular. These systems are normally designed for improving automatic or autonomous active mobility of humans, increasing their safety, situational awareness and real time perception of the environment. A variety of solutions have been developed, such as positional navigation aids based on sensors, objects detection systems in visual perception, portable devices or intelligent sound/haptic feedbacks. We then present a survey of the most promising and recent related work in vision-driven assistive systems, real-time detection of objects, pedestrian analysis and understanding footpath scenes as well as wearable navigation aids, multimodal interaction techniques.

Visually impaired and blind (VIB) assistive technologies have made great strides due to computing vision, deep learning, and edge computing. Several conventional navigation aids including white cane, guide dog, ultrasonic sensor, RFID-based system and GPS devices have been utilized for mobility and obstacle avoidance. These methods increase the superficial safety but are not able to truly understand general environments semantically, resulting impracticable particularly in dynamic indoor/outdoor situations [19], [20]. The vision-based systems enabled a significant transition towards intelligent assistive technologies. In one of the earlier camera-based systems, hand-designed features and classical machine learning were exploited that did not handle well variations like illumination changes, occlusion and cluttered background. The success of the convolutional neural networks (CNN) significantly boosted object detection and recognition accuracy and opened the door for real-time perception based assistive navigation applications [21].

The YOLO-based object detection models became popular for two reasons: one-stage structure and real-time inference. A YOLO based IOT object recognition for the visually impaired with audio feedback was introduced by Rahman and Sadi [22]. Hussan et al. proposed a real-time object detection using yolov3, but the performance of its model was not precise and computational consumption was too heavy [23]. Islam et al. used SSDLite MobileNetV2 on a head-mounted assistive device, offering fair performance but being not seamlessly mobile-integrated [24]. Kumar and Jain developed a YOLO-powered navigation system with smart stick that achieved 89.24% accuracy, but still failed to generalize well to especially challenging real-world scenarios [25]. In recent works, researchers tried to achieve a better trade-off between accuracy and efficiency by introducing lightweight and engineered detection models. Gabriel et al. deployed YOLOv8 on Raspberry Pi for real-time object recognition, achieving enhanced accuracy with some latency issue regard- ing the edge devices [26]. Alahmadi et al. improved YOLOv4 with ResNet101 backbone to enable better feature learning on assistive navigation with increased computation [27]. Arifando et al. proposed a lightweight model based on the YOLOv5 architecture, which integrated GhostConv and C3Ghost modules to significantly decrease FLOPs but with detection accuracy remain [28].

The safety of footpath and pedestrian walkways is becoming an important trend in outdoor navigation research. PFPN-ADT-based anomaly detection model to detect vehicles and abnormal objects in pedestrian walkways using also Sophia and Chitra [29]. Pustokhina et al. proposed a DLADT-PW model, where the deep learning method is Mask R-CNN with DenseNet-169, in order to automate the process and achieved high accuracy for UCSD dataset at cost of the significant computational resources [30]. Alohali et al. proposed a federated learning based anomaly detection model (ADPW-FLHHO) with remote sensing imagery without a privacy issue to some extent, but it cannot be deployed in real time on edge devices [31]. Morra et al. inquired about sidewalk accessibility mapping with smartphone image and visual AI but not leveraged for real-time edge intelligence [32].

Edge intelligence and wearable assistive systems have been explored to improve portability and responsiveness. Mahendran et al. proposed a mobile edge AI-based assistive system using SSD MobileNet and depth sensors, but the backpack-based hardware reduced usability [33]. V. K. et al. implemented a navigation assistance based on Raspberry Pi with MobileNet SSD and ultrasonic sensors, however an accuracy of 72% was obtained due to scarce training data [34]. Lima et al. proposed a smartphone- based positioning aid scheme which employs ORB and KNN with little processing time cost but inferior robustness than deep-learning-based approaches [35]. Smart glass technology offers an eye-free interface and has drawn attention. Daescu et al. proposed a smart-glass-based face recognition system based on dcnn, while frequent retraining was needed for adapting the new identities [36]. Chen et al. proposed a cloud-based wearable object recogni-

tion system via low-cost devices, with emphasis on reducing device side computations and introducing dependence on network connectivity [37]. Chang et al. pioneered obstacle avoidance, fall detection and zebra-crossing assistance systems based on smart glass, which can enhance safety (but with small scalability [38], [39]).

More recent work focuses on optimal and generalized assistance schedulers. Kadam et al. propose an edge-based dangerous object detection system based on small-sized deep learning models [40]. Bhandari et al. used transfer learning with YOLOv8 to increase the adaptability to new objects for visually impaired performers [41]. Vision–language and multimodal assistive devices e.g., VocalEyes combined object detection with text-to-speech and context awareness, leading to situational awareness but computationally complexity [42]. In general, despite that the state-of-the- art can achieve promising results in object detection, footpath analysis, smart wearables and multimodal interaction as point out earlier on existing ATs, most of systems are domain specific, computationally intensive and data-driven at feature level. These drawbacks are the main motivation behind proposing a lightweight, unified and real-time multifeature assistive system—which directly relates to what is proposed in [43], [44], [45].

Although significant achievements have been made in current assistive systems, some limitations have not been addressed yet. Most of the previous solutions focused on separate tasks such as the detection of objects, footpath analysis, face recognition or text-based recognition without providing a general perception framework. Most of the existing systems use computationally expensive model or fixed datasets, which hinder them from being deployed on resource-constrained edge devices and adapted to real environment. In addition, the combination of multi-feature visual perception with real-time auditory feedback in a lightweight framework has not been well addressed. These issues lead to the necessity of a seamless, online and optimization-based multi-feature assistive system that combines object detection, footpath safety Analysis. face familiarity recognition, OCR (Optical Character Recognition) and intelligent voice assistance – aiming at both accuracy considerations and practical utility for visually impaired participants.

III. METHODOLOGY

In this section, we describe the method of our proposed deep learning based assistive navigation system to improve environmental perception for visually impaired users. The system is based on a multi-feature detection model that integrates the real-time moving object detection, recognizing Bangladesh currency, footpath security monitoring, known and unknown face identification and Bangla–English character recognition in optical character recognition (OCR) under a single framework. An embedded real-time voice assistant is also used to provide an intuitive and affordable interaction by transforming the perception results into user-friendly audio presentation. The general goal is to ensure in-time, contextual-aware, and eyes-free assistance for promoting safer mobility and situational awareness within complex scenarios.

Fig. 5 shows the overview of the implementation system. The workflow consists of an end-to-end pipeline which comprises image capturing and dataset generation, annotation, preprocessing, deep learning-based model developing. In the testing phase, live streaming video is uploaded by a camera and analyzed by pre-trained detection models and supplementary vision modules for perception of semantic context. The integrated results of these modules allow the system to detect obstacles, recognize currency notes, measure sidewalk safety (open space), differentiate familiar and unfamiliar people, and understand Bangla text information corresponding to assist the blind community in the daily task of assistance for walking. The combined output from these perception modules is also given as input to a voice assistant module, which uses it to attend important information and provides the end users (blind people) with auditory feedback for safe and independent traversing.

A. DATASET DESCRIPTION

A balanced and representative dataset is an important requisite for training reliable deep learning models, even more in the case of assistive vision systems where a wrong prediction can have a direct impact on user safety. For implementing the multi-feature detection module of the novel navigation aid (please refer to Figure 1), three different datasets were used: i.e., an environmental object detection dataset, a Bangladesh currency image(s) detection data set and; a blockage print (of stairs)/plain floor tripping risk prediction dataset. More importantly, the two datasets support this system to provide navigation aid, transaction and safety operation for the visually impaired persons.

1) Object Detection Custom Dataset

A custom object detection dataset was prepared on which to train the environment tracking module of our platform. The **703 raw and unique images** were acquired from various indoor and outdoor environments in Pabna University of Science and Technology (PUST), Bangladesh. The data were collected in academic buildings, access roads, internal roads, pedestrian paths, green areas and lake area. To simulate the real navigation, images of vehicles were also collected at nearby traffic intersections adjacent to the university highway. The dataset contain **nine object categories**: Vehicle, Chair, Door, Man, Road, Stair, Table, Tree and Wall. As there are multiple object categories in the same image, we get **812 class-wise image instances** although there remains 703 unique images. This leads to an excess of annotated object instances compared to the number of images. After verifying the annotation and pre-processing, the dataset was split into **563 training**, **71 validation**, and **69 test images**. All data augmentation was applied to only the training portion of the data and resulted in **1829** images for training with both validation and test sets maintained as is for a fair testing.

TABLE 1. Comparison of Existing Vision-Based Assistive Systems for Visually Impaired Users

Ref.	Year	Dataset	Method	Key Contribution	Limitations
[22]	2021	Custom	YOLO	IoT-based object recognition with audio feedback	Limited scalability
[23]	2022	Custom	YOLOv3	Real-time object detection	Low precision, heavy model
[24]	2023	COCO	SSDLite MobileNetV2	Head-mounted assistive system	Not fully mobile, moderate accuracy
[25]	2022	Custom	YOLO + Smart Stick	Path detection and navigation assistance	Limited adaptability
[26]	2025	COCO	YOLOv8	Edge-based object recognition	Latency on edge devices
[27]	2023	Custom	YOLOv4-ResNet101	Improved feature extraction	High computational cost
[28]	2023	Custom	YOLOv5 + GhostConv	Lightweight detection with reduced FLOPs	Single-task focus
[29]	2023	UCSD	PFPN-ADT	Pedestrian walkway anomaly detection	No real-time deployment
[30]	2021	UCSD	Mask R-CNN	High-accuracy anomaly detection	Heavy model
[31]	2023	UCSD	ADPW-FLHHO	Federated learning-based anomaly detection	Limited edge usability
[32]	2024	Crowd-sourced	PSPUNet	Sidewalk accessibility mapping	Not optimized for edge AI
[33]	2021	Multi	SSD MobileNet	Mobile edge AI-based navigation	Bulky wearable hardware
[34]	2023	Custom	MobileNet SSD	Low-cost navigation aid	Small dataset
[35]	2023	Custom	ORB + KNN	Smartphone positioning assistance	Less robust than DL models
[36]	2019	Custom	CNN-based Face Recognition	Smart glass-based face identification	Requires frequent retraining
[37]	2019	Custom	Cloud-based DL	Low-cost wearable recognition	Network dependency
[38]	2021	Custom	DL + Edge Computing	Zebra-crossing safety assistance	Scenario-specific
[39]	2020	Custom	Smart Glass + DL	Obstacle avoidance and fall detection	Device dependency
[40]	2025	Custom	Lightweight DL	Edge-based hazardous object detection	No multi-feature fusion
[41]	2024	Custom	YOLOv8 + Transfer Learning	Adaptation to new objects	Dataset-dependent
[42]	2023	Custom	Vision–Language Model	Context-aware assistive feedback	High computational cost
[43]	2025	Custom	DL-based Navigation	AI-based navigation assistance	Limited real-time validation
[44]	2020	Custom	Multi-object DL	Smart navigation framework	Lacks edge optimization

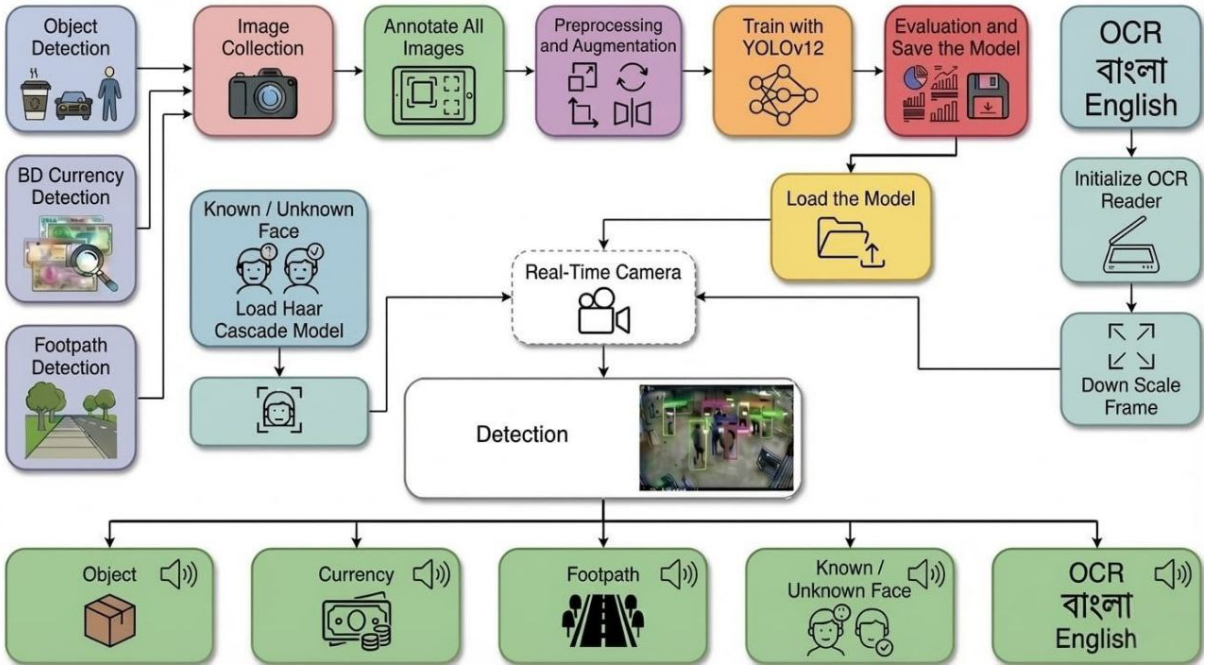

FIGURE 1. Overall architecture of the proposed multi-feature detection system

2) Bangladesh Currency Detection Dataset

In order to help the currency recognition module a Bangladesh currency dataset was established with samples that were collected from PUST's students, teachers and officers. The currency notes and coins were captured in photographs of individual bank note/coin by mobile phone camera to represent the real working environment as blind people will be using it for their financial transactions on daily ba-

FIGURE 2. Simple image for object detection dataset

FIGURE 3. Simple image for BD currency detection dataset

sis. The dataset consists of 10 currency classes that include: banknotes with denominations (2,5 10, 20, 50, 200, 500,and 1000)taka, as well as coins worth (1, 2 and 5) taka. Following preprocessing and quality control, we retained **1627** images. In this work, we split the dataset into **1270 training images**, **168 validation images** and **189 test images**. Data augmentation was performed only on the training set to augment the number of training samples up to **3801**. The images present different environment settings (e.g., full of sun shine, indoors with strong shadow, mixed perspectives) and backgrounds containing tables, clothes, desks or public scenes. More hard examples like folded notes, worn surfaces and texture changes will also benefit its realism.

3) Footpath Detection Dataset

The footpath detection of the proposed system uses a pre-collected public dataset supplied from the Roboflow platform. Unlike previous datasets, this dataset was directly borrowed (i.e., not handcrafted) from its original authors. It is composed of images split into **four** footpath accessibility classes: safe, occupy fullly, hide occupily, and not-safe. The dataset, containing **4884 images** was pre-segmented to contain **3914 training samples**, **735 validation samples** and finally **235 test samples**. These images are from real-world pedestrian scenes, including dense lane-way, partially obscured path, dangerous area and open sidewalk. All annotations are written in YOLO format. This dataset is important for training and evaluating models which can recognize pedestrian safety and warn visually-impaired persons against the potential hazards when traveling.

FIGURE 4. Simple image for footpath detection dataset

B. IMAGE ANNOTATION

Accurate annotation to help train robust deep learning models is vital, it's a particularly crucial factor in assistive technologies where misclassification can lead to unacceptable performance in terms of user safety. Consequently, extensive manual annotation is done for both object detection dataset and Bangladesh currency dataset.

TABLE 2. Class-wise Distribution of the Datasets Used in the Proposed Framework

Dataset	Class	Images	Instances
Object Detection	Vehicle	90	145
	Chair	90	125
	Door	90	112
	Man	91	125
	Road	90	95
	Stair	90	94
	Table	91	101
	Tree	90	138
	Wall	90	109
Currency Detection	Five Hundred Taka	153	297
	Fifty Taka	226	472
	Five Taka	199	629
	One Taka	88	292
	One Thousand Taka	126	334
	Ten Taka	205	502
	Twenty Taka	183	553
	Two Hundred Taka	146	169
	One Hundred Taka	229	584
	Two Taka	223	464
Footpath Detection	free for use	2345	2373
	Fully Occupied	1413	1416
	Not for Safe	47	47
	Partially Occupied	1079	1082

1) Annotation of Object Detection Dataset

All object detection images were copied to Roboflow for labelling. Each instance of the objects was annotated manually in YOLO bounding box format which represents the class label and normalised coordinates of a bounding box. Precise annotations of the bounding boxes helped in detecting objects (doors, staircases, vehicles and pedestrians) important for obstacle avoidance.

To ensure consistency across all nine object classes, Roboflow's annotation tools were employed. Especially overlapping objects and size differences between the blobs were subject to careful setting of spatial borders. Hand checking was performed on annotated images, to clean label errors and ensure the quality of the dataset.

2) Annotation of Bangladesh Currency Dataset

Money notes and coins were annotated with the roboflow platform. The denomination of each sample was identified by bounding boxes around the sum number and specific image pattern or icon of the note / coin. It is important for annotation to be precise in currency recognition, since small errors in placement of annotations can result in large influences on the classification performance.

A wide range of currency images (clean, worn, folded or partially shadowed samples) were analyzed intensely to provide accurate annotation quality. Class balance was verified by Roboflow's audit tools to ensure that each denomination was well-represented.

C. PREPROCESSING AND DATA AUGMENTATION

To improve model robustness and generalisability, preprocessing and data augmentation were implemented for object detection and Bangladesh currency datasets.

1) Preprocessing

Both datasets were processed with a series of preprocessing steps on Roboflow:

- Auto-orient based on meta data to correct image orientation.
- Enhance the contrast to see the images clearer under low light condition.
- Colour and exposure norming for consistent model input.
- Consistent file renaming and systematic organization of datasets for reproducibility.

Preprocessing images such as these put the same image values recorded under varying conditions to the unique neural network input data range, and enhanced the descriptiveness of features in a feature map obtained from the neural network.

2) Data Augmentation

To have a diversity of images in the dataset, and to avoid gathering even more training images, several augmentation techniques were used:

- Horizontal reflection to generate mirrored perspectives.
- Random rotation between -10° and $+10^\circ$ to simulate camera tilt.
- Brightness change ($\pm 15\%$) for indoor and outdoor.
- Simulatemotion blur and defocus effects using gaussian blur.
- 0.1% of pixels corrupted by noise to simulate sensor noise injection.

These augmentation techniques allow the model to learn invariant feature representations and to have reasonable performance under realistic scenarios including non-uniform illumination, shading, and motion artifacts.

D. TRAINING WITH YOLOV12

For the perceived scenes are complicated and timely, YOLOv12 is taken as the basic deep-learning network in OCR application for object detection, recognition of currency and classification of footpath. YOLOv12 is a further step in real-time object detection thanks to introducing attention-driven architectures that are focused on both computation efficiency and feature representation quality. Based on the demonstrated successes of previous YOLO versions then, our key focus in designing YOLOv12 is on architectural tweaks that further maximise detection accuracy and specifically preserving real-time inference performance characteristics which is crucial for applications such as assistive navigation system for blind persons.

The re-design of the feature extraction is at the heart of YOLOv2, which consists Residual Efficient Layer Aggregation Network (R-ELAN), FlashAttention, and 7×7 separable

convolutions as visualized in Fig. 4 [46]. Together, these components reinforce feature discrimination and better spatial knowledge while keeping the computational burden low so that our model can work efficiently even under visually cluttered and dynamically varying environments.

One characteristic of YOLOv12 is that it has good adaptability to difficult detection scenarios. The use of area-based attention mechanism, which is sped up by FlashAttention, enables the network to focus on informative spatial regions and ignore irrelevant background features. This ability considerably enhances the localization of small, partly occluded, and overlapping objects which are often present in realistic navigation settings. Although the representational augmentation provided by attention mechanisms, performance-wise YOLOv12 manages to retain the high speed of the YOLO family of detectors, which shows that it can be useful for tasks sensitive to time and autonomous perception and assistive perception [46].

Aside from the navigation-centric tasks, YOLOv12 also shows good flexibility in a wide range of applications. Its superior detection capability – designed specifically for smart transportation systems (Smart TSS) is contributed to the reliability of ADAS features, including vehicle/pedestrian detection and traffic sign recognition. For example, in the fields of health care and agriculture surveillance, the ability of YOLOv12 to detect small or visually difficult objects with limited computational overhead underscores its potential in real-world applications. Such properties enable YOLOv12 easy to be employed as a backbone for the multi-feature detection pipeline.

The YOLOv12 network adopts an integrated end-to-end detection framework where the single network simultaneously regresses bounding box together and classifies object. The architecture consists of three main parts: the backbone, which is used to extract multi-scale features representation; neck, which gathers and retouches these features; and head part, that produces the final detection results.

1) Backbone

The backbone of YOLOv12 is responsible for transforming raw input images into rich and discriminative multi-scale feature representations. Central to this process is the Residual Efficient Layer Aggregation Network (R-ELAN), which combines deep convolutional layers with strategically placed residual connections. This design mitigates gradient degradation and enhances feature reuse, allowing the network to capture fine-grained spatial details across objects of varying scales.

Compared to earlier YOLO variants, YOLOv12 introduces lightweight convolutional blocks that prioritise parallelisation and reduced computational complexity. These operations can be generically expressed as:

$$F_{\text{out}} = \sum_{i=1}^{n} W_i * F_{\text{in}} + b_i \tag{1}$$

where F_{out} denotes the output feature map, W_i represents convolutional filters, F_{in} is the input feature map, and b_i denotes the bias term. By distributing computations across multiple smaller convolutions rather than relying on large kernels, YOLOv12 achieves faster processing while maintaining strong feature extraction capability.

To further enhance efficiency, the backbone employs 7×7 separable convolutions, which preserve spatial context while significantly reducing parameter count and floating-point operations. In addition, multi-scale feature pyramids are constructed to ensure effective representation of objects with diverse sizes and aspect ratios, including small or partially occluded targets.

2) Neck

The neck component acts as an intermediary between the backbone and the detection head, aggregating and refining multi-scale feature representations. YOLOv12 integrates an area-based attention mechanism within the neck, accelerated using FlashAttention, to enhance the model's focus on salient regions in cluttered scenes. This attention mechanism can be mathematically formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{2}$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d_k represents the dimensionality of the key vectors. By applying attention at the area level rather than at individual pixel locations, YOLOv12 reduces memory overhead and enables efficient processing of high-resolution feature maps.

3) Head

The detection head of YOLOv12 converts the refined feature representations into final bounding box predictions and class probabilities. The head employs streamlined multi-scale detection branches and refined loss formulations to balance localisation accuracy and classification confidence. The overall training objective can be expressed as:

$$\mathcal{L} = \lambda_{\text{coord}} \sum (\hat{x}-x)^2 + (\hat{y}-y)^2 + \lambda_{\text{obj}} \sum (\hat{C}-C)^2 + \cdots \tag{3}$$

where $\hat{x}$ and $\hat{y}$ denote the predicted bounding box coordinates and $\hat{C}$ represents the object confidence score. This multi-component loss formulation enables robust optimisation of detection performance across diverse environmental conditions while maintaining real-time inference capability.

4) Training Configuration

YOLOv12 was trained using the Ultralytics framework with a set of carefully selected hyperparameters to ensure a balance between detection accuracy and real-time performance. Training was conducted for 100 epochs using stochastic gradient descent (SGD) optimisation with an initial learning rate of 0.01 and a momentum value of 0.937. A batch size of 16 and a fixed input image resolution of 640×640 pixels

were employed throughout training. The model comprises 272 layers with approximately 2.6 million trainable parameters and requires 6.7 GFLOPs, making it suitable for deployment in assistive navigation systems operating under computational constraints. The complete training configuration is summarised in Table 3.

TABLE 3. Training Parameters for YOLOv12

Parameter	Value
Epochs	100
Batch Size	16
Input Image Size	640×640
Optimizer	SGD
Initial Learning Rate	0.01
Momentum	0.937
Number of Layers	272
Number of Parameters	2,568,828
GFLOPs	6.5

The selected training configuration reflects a deliberate trade-off between computational efficiency and detection accuracy. The relatively low parameter count and GFLOPs, combined with the use of SGD optimisation, enable stable convergence and real-time inference, making YOLOv12 particularly well suited for the proposed deep learning-powered navigation aid for visually impaired users.

E. KNOWN AND UNKNOWN FACE DETECTION

As part of the proposed deep learning-powered navigation aid, recognising familiar and unfamiliar individuals in the surrounding environment is crucial for enhancing social awareness and interaction for visually impaired users. In everyday navigation scenarios, identifying whether a nearby person is known or unknown allows the system to provide meaningful contextual information, thereby improving user confidence and situational understanding. To fulfil this objective, a real-time known and unknown face detection process is integrated into the multi-feature assistive framework.

The face recognition pipeline operates in a lightweight yet effective manner to ensure real-time performance on resource-constrained devices. It begins with face localisation from live camera frames, followed by preprocessing, similarity-based identity matching, and a threshold-driven decision mechanism. This design choice prioritises computational efficiency while maintaining reliable recognition under practical operating conditions.

Let an input frame captured by the camera be represented as

$$I \in \mathbb{R}^{H \times W \times 3} \tag{4}$$

To reduce computational overhead, the frame is first converted into a grayscale image:

$$I_g = \text{Grayscale}(I) \tag{5}$$

Face regions are then detected using a pre-trained Haar Cascade frontal face detector provided by OpenCV, which is based on the boosted cascade framework proposed by Viola and Jones [47], [48]. The Haar Cascade classifier applies a sequence of weak learners using Haar-like features to rapidly scan the image. For a given sliding window w, the classifier output can be expressed as:

$$C(w) = \begin{cases} 1, & \text{if a face is detected,} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Once a face is detected, the corresponding region of interest is extracted and resized to a fixed resolution of 100×100 pixels to ensure uniformity across all samples:

$$F_t \in \mathbb{R}^{100 \times 100} \tag{7}$$

A database of known individuals is constructed offline using labeled facial images. Each stored face undergoes the same preprocessing steps as the real-time detected faces. The known face database can be defined as:

$$\mathcal{D} = \{(F_k, y_k)\}_{k=1}^{N} \tag{8}$$

where F_k represents the k-th stored grayscale face template and y_k denotes the corresponding identity label.

To determine whether a detected face belongs to a known individual, the system computes a similarity score between the detected face F_t and each stored template F_k. Normalised cross-correlation is used for similarity measurement, implemented via OpenCV template matching. The similarity score is defined as:

$$S(F_t, F_k) = \frac{\sum (F_t - \bar{F}_t)(F_k - \bar{F}_k)}{\sqrt{\sum (F_t - \bar{F}_t)^2 \sum (F_k - \bar{F}_k)^2}} \tag{9}$$

where $\bar{F}_t$ and $\bar{F}_k$ denote the mean pixel intensities of the detected and stored face templates, respectively. The similarity score lies in the range $[-1, 1]$, with higher values indicating stronger resemblance.

The maximum similarity score across all stored identities is computed as:

$$S_{\max} = \max_k S(F_t, F_k) \tag{10}$$

The final identity decision follows a threshold-based open-set recognition strategy [49]. A detected face is classified as known if the maximum similarity score exceeds a predefined threshold τ; otherwise, it is labelled as unknown:

$$\text{Identity}(F_t) = \begin{cases} \text{Known}(y_k), & \text{if } S_{\max} \geq \tau, \\ \text{Unknown}, & \text{otherwise.} \end{cases} \tag{11}$$

In this work, the threshold is empirically set to $\tau = 0.6$, which provides a balance between false acceptance and false rejection in real-world conditions. Known faces are highlighted with green bounding boxes, while unknown faces are marked in red, enabling immediate and intuitive feedback for the assistive navigation system.

All detection events are logged with timestamps and confidence scores for further analysis and system evaluation. Although deep learning-based face recognition models can achieve higher accuracy, the proposed classical approach offers a favourable trade-off between performance and computational efficiency. This makes it well suited for real-time

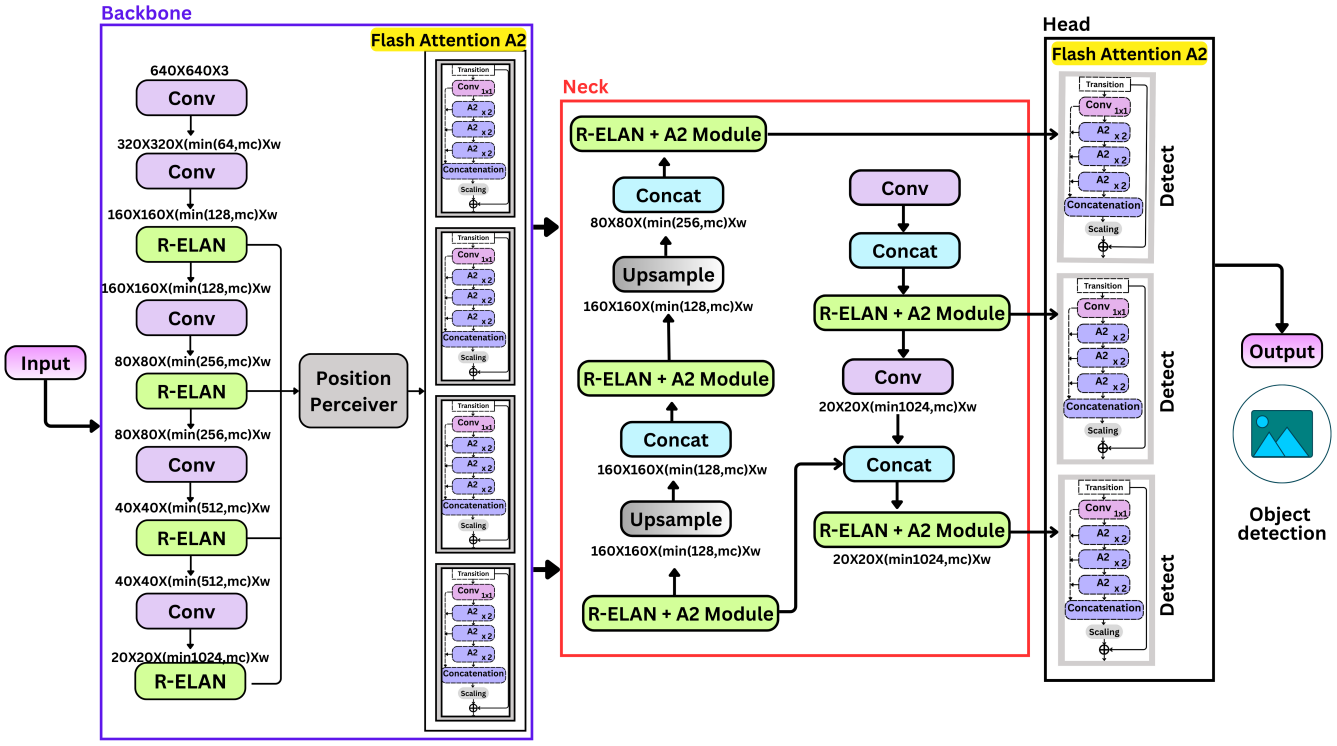

FIGURE 5. YOLOv12 architecture

assistive navigation systems designed for visually impaired users, where responsiveness and reliability are of paramount importance.

F. OPTICAL CHARACTER RECOGNITION

Textual information available in the surrounding environment, such as signboards, room labels, warning notices, and public instructions, plays a crucial role in safe navigation and situational awareness. For visually impaired users, the ability to recognise such text in real time can significantly enhance independence and confidence during daily mobility. To address this requirement, the proposed system incorporates a real-time Optical Character Recognition (OCR) module capable of recognising both Bangla and English text from live video streams.

The OCR module is implemented using EasyOCR, a deep learning-based multilingual text recognition framework that combines convolutional neural networks (CNNs) for visual feature extraction and recurrent neural networks (RNNs) for sequence modelling [50]. EasyOCR has demonstrated strong performance across multiple scripts, including Indic languages, and is particularly suitable for Bangla text recognition due to its ability to handle complex character structures, ligatures, and font variations [51], [52].

Let an input image frame captured by the camera be represented as:

$$I \in \mathbb{R}^{H \times W \times 3} \tag{12}$$

To ensure real-time performance, the frame is resized before OCR processing:

$$I_r = \mathcal{R}(I) \tag{13}$$

where $\mathcal{R}(\cdot)$ denotes image resizing to a fixed resolution. The OCR process begins with text region detection, where candidate text areas are localised within the image. For each detected region, a cropped text patch is extracted as:

$$T_i = I_r[x_i : x_i + w_i,\ y_i : y_i + h_i] \tag{14}$$

Each text patch is then passed through a convolutional feature extractor to learn discriminative visual representations:

$$F_i = \mathrm{CNN}(T_i) \tag{15}$$

followed by sequence modelling using a recurrent neural network to capture contextual dependencies between characters:

$$H_i = \mathrm{RNN}(F_i) \tag{16}$$

This sequence-based modelling is especially important for Bangla text, where character shapes and meanings are often context-dependent.

The final recognised text is obtained using a Connectionist Temporal Classification (CTC) decoding strategy, which enables character sequence prediction without requiring explicit character-level segmentation:

$$\hat{Y}_i = \mathrm{CTC\_Decode}(H_i) \tag{17}$$

The OCR output for a given frame can therefore be expressed as:

$$\mathcal{O}_{\text{OCR}} = \{\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_N\} \tag{18}$$

where N denotes the number of detected text regions.

To ensure correct visualisation of Bangla characters, which require proper Unicode shaping and glyph rendering, the recognised text is displayed using Bengali-compatible TrueType fonts. During real-time operation, the OCR module runs in parallel with object detection, currency recognition, face detection, and footpath analysis modules. Although transformer-based OCR models have recently achieved higher recognition accuracy [53], [54], the selected EasyOCR-based approach offers a favourable trade-off between accuracy, computational efficiency, and real-time performance. This makes it particularly suitable for assistive navigation systems deployed on resource-constrained platforms.

The complete workflow of the process is shown Algorithm 1 of the envisioned multi-feature deep learning–enabled navigation aid for a visually impaired individual. The algorithm starts by loading up all the pre-trained models that are to be used, these include YOLOv12 for object, money and footpath detection; Haar Cascade classifier for face detection and the EasyOCR engine for multilingual text recognition. During time of execution, live video frames are acquired in real-time from the camera and pre-processed by resizing and normalization to keep consistent input quality. At which every frame is fed to YOLOv12 model for the detection of environmental ojects, currency denomination recognition and footpath safety condition analysis simultaneously. Simultaneously to this, the system does face detection by converting the frame to grayscale and applying a Haar Cascade detector. The detected faces are matched based on similarity with the database to decide whether a face is known or unknown. OCR is applied later to recognize Bangla and English text information in the scene. At last, the outputs of all perception modules fuse to get the context and situation scene-aware feedback. The processed information is then converted into a real-time audio or textualap guidance for safe and independent navigation of visually impaired individuals through dynamic environments.

Algorithm 1 Multi-Feature Deep Learning–Powered Navigation Aid for Visually Impaired

Require: Live video stream V from camera
Ensure: Real-time auditory/visual feedback for navigation assistance

Load pre-trained YOLOv12 models for object detection, currency detection, and footpath classification
Load Haar Cascade face detector
Load known face database $\mathcal{D}$
Load EasyOCR model for Bangla and English text
while camera is active **do**
    Capture frame I_t from video stream V
    Preprocessing:
    Resize I_t to 640×640
    Normalize pixel values
    Object, Currency, and Footpath Detection:
    $\mathcal{O}_t \leftarrow \text{YOLOv12}(I_t)$
    Extract detected objects, currency denominations, and footpath safety class
    Face Detection and Recognition:
    Convert frame to grayscale I_g
    Detect faces using Haar Cascade
    for each detected face F_t **do**
        Resize F_t to 100×100
        Compute similarity score $S(F_t, F_k)$ for each $F_k \in \mathcal{D}$
        $S_{\max} \leftarrow \max(S)$
        if $S_{\max} \geq \tau$ **then**
            Label face as **Known**
        else
            Label face as **Unknown**
        end if
    end for
    Optical Character Recognition (OCR):
    Detect text regions $\{T_i\}$ in I_t
    for each text region T_i **do**
        Recognize text using EasyOCR
    end for
    Decision Fusion:
    Convert fused perception results into voice instructions using TTS
    Deliver real-time auditory feedback via the voice assistant
end while

IV. RESULT AND DISCUSSION

The proposed YOLO-based multi-feature assistive framework was tested using the testing datasets after the model was trained. System The performance of the system was measured using some standard evaluation measures such as Precision, Recall and mean Average Precision (mAP@0.5 and mAP@0.5-0.95) Recognition Accuracy, False Rejection Rate (FRR), False Acceptance Rate (FAR), Character Error Rate (CER), Word Error Rate(WER) and Word Recognition Accuracy. The metrics were calculated using the confusion matrix of TP (true positives), TN (true negatives) FP(false positives), FN(false negatives) to thoroughly evaluate detection accuracy, recognition reliability and real-time applicability.

For object detection, footpath detection and Bangladesh currency recognition, the main evaluation metrices were Precision, Recall and mAP. Precision refers to the ability of a model not to label as positive a sample that is negative and Recall measures the capacity of a model to find all positive samples. The mean average precision (mAP) scores confidence of classification and accuracy of localisation for all classes, so mAP is a strong measure of the detection performance overall.

- **Average Precision by Class (mAP):** Mean average pre-

cision (mAP) is one of the more commonly applied metrics for object detection problems. The mean Average Precision (AP) is computed using the number of True Positives (TP), False Positives (FP), and False Negatives (FN) for each class and then the aggregate mean across all classes.

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{19}$$

where N denotes the total number of classes and AP_i represents the Average Precision of the i-th class, computed as the area under the Precision–Recall (PR) curve.

- **Precision:** Precision quantifies the proportion of correctly identified positive instances among all instances predicted as positive. It is computed by dividing the number of true positives (TP) by the sum of true positives and false positives (FP), as shown in Equation 20. This metric reflects the model's ability to avoid false alarms in classification tasks

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{20}$$

- **Recall:** Recall measures the model's ability to correctly identify all relevant positive instances. It is calculated by dividing the TP by the sum of true positives and false negatives (FN), as expressed in Equation 21. A higher recall indicates greater sensitivity to actual target objects, minimizing missed detections

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{21}$$

The recognition performance for the known and unknown face identification module was assessed in terms of Recognition Accuracy, Recall, FRR (False Rejection Rate) andFAR (False Acceptance Rate).

- **Recognition Accuracy:** Recognition Accuracy is the human performance measure of how properly a face recognition system works, and represent what proportion of true classed known and unknown faces were able to be correctly classified. This measure expresses the system's capacity to accurately identify subjects without misclassifying them.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{22}$$

- **False Rejection Rate (FRR):** FRR is the ratio of missing identities amongst known samples. A low FRR is especially important in assistive systems so that users do not become confused or frustrated by having to respond again and again to identity queries.

$$\text{FRR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{23}$$

- **False Acceptance Rate (FAR):** FAR measures the probability of erroneously classifying a stranger as an enrollee. It is essential to minimize FAR in order to guarantee security and prevent deceptive identity feedback.

$$\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{24}$$

- **Character Error Rate (CER):** CER calculates the number of character-level transcription errors in the OCR output by normalizing the edit distance between predicted and ground-truth text. It is especially valuable for the assessment multilingual OCR results.

$$\text{CER} = \frac{\text{Edit Distance}}{\text{Total Characters}} \tag{25}$$

- **Word Error Rate (WER):** WER is a word-level OCR performance metric that calculates the insertion, deletion and substitution errors on the recognized and reference texts. A smaller WER indicates better word-level recognition accuracy.

$$\text{WER} = \frac{\text{Edit Distance}}{\text{Total Words}} \tag{26}$$

- **Word Recognition Accuracy:** Word Recognition Accuracy (WRA) is the percentage of correctly recognized words and thus serves as a natural indicator of OCR readibility in assistenc technology.

$$\text{Word Accuracy} = (1 - \text{WER}) \times 100 \tag{27}$$

The performance on the Bangladesh currency detection dataset is presented in Table 4. High level, our model has an overall precision at 0.922, recall at 0.931 and mAP@0.5 of 0.951, which reflects the high detection accuracy for various amount of currencies. For high valued notes also One Thousand Taka, One Hundred Taka and Ten Taka mapped the particularly highest mAP@0.5 are familiar with the visual distinctive attributions and the relatively bigger volume. Lower mAP@0.5-0.95 for lower denominations like Two Taka and Five Taka indicating that similar looking notes, dirty notes with less texture due to the ageing in chests and strong lighting is common to all denomination of notes. However, the results overall show the proposed system to be reliable for real-world currency recognition allowing visually impaired individuals to transact financially more independently.

Table 5 lists the results of our approach on footpath detection. Finally, the overall mAP@0.5 of the methods is produced by our model. then its ability to interpret pedestrian walkway conditions can now be seen. The "Free for Use" class gets high values of precision (0.808) and recall (0.945), demonstrating that the model is able to make reliable predictions on safe walking paths, which are essential for assistive navigation. Comparatively, 'Fully Occupied' and 'Not Safe for Use' classes performs worse due to class imbalance, reduced sample size or visual occlusion in challenging outdoor surroundings. Despite these difficulties, the recall values obtained are acceptable and will enable the system to raise timely alerts for non-safe walking conditions and eventually help in user safety.

The class-wise results of the object detection dataset are reported in Table 6. The overall mAP@0.5 is obtained from the proposed framework of 0.625 for all object classes. For example, Door, Road as well as Chair and Stair achieve relatively higher detection rate since they have steady geometrical

TABLE 4. Performance Evaluation of the Bangladesh Currency Detection Dataset

Class Name	Images	Instances	Precision	Recall	mAP@0.5	mAP@0.5–0.95
All Classes	168	433	0.922	0.931	0.951	0.584
Five Hundred Taka	38	68	0.852	0.824	0.902	0.392
Fifty Taka	18	39	0.946	0.923	0.929	0.428
Five Taka	19	60	0.884	0.850	0.911	0.470
One Taka	9	26	0.861	0.962	0.966	0.747
One Thousand Taka	14	36	0.911	0.972	0.979	0.710
Ten Taka	21	48	1.000	0.969	0.984	0.581
Twenty Taka	14	41	0.968	0.951	0.962	0.563
Two Hundred Taka	37	37	0.963	1.000	0.979	0.915
One Hundred Taka	15	36	0.964	1.000	0.984	0.656
Two Taka	20	42	0.871	0.857	0.911	0.378

TABLE 5. Performance Analysis of the Footpath Detection Dataset

Class Name	Images	Instances	Precision (P)	Recall (R)	mAP@0.5	mAP@0.5–0.95
All Classes	735	736	0.653	0.700	0.666	0.520
Free for Use	594	595	0.808	0.945	0.914	0.752
Fully Occupied	19	19	1.000	0.381	0.598	0.535
Not Safe for Use	12	12	0.405	0.682	0.610	0.418
Partially Occupied	110	110	0.397	0.791	0.541	0.374

TABLE 6. Class-wise Performance of the Object Detection Dataset

Class Name	Images	Instances	Precision (P)	Recall (R)	mAP@0.5	mAP@0.5–0.95
All Classes	71	126	0.692	0.554	0.625	0.472
Vehicle	16	19	0.870	0.421	0.471	0.253
Chair	17	24	0.932	0.571	0.823	0.657
Door	12	17	0.755	0.882	0.894	0.787
Man	6	10	0.597	0.600	0.688	0.456
Road	12	12	0.955	0.583	0.807	0.676
Stair	8	8	0.632	0.750	0.799	0.522
Table	6	8	0.511	0.875	0.555	0.462
Tree	5	9	0.336	0.111	0.292	0.226
Wall	12	19	0.642	0.191	0.293	0.207

configuration and confident local context relevance in the dataset. Performance is lower for the Wall and Tree classes, primarily owing to scale variation, background clutter, and limited training examples. It worth noting that door class gets an excellent Recall value, which is especially crucial in the assistive navigation system because failing to notice these obstacles may have severe safety implications. These results show that although object detection is a remain challenging task, the presented model offer stable and reliable performance for navigating essential objects.

The computational complexity and real time of the proposed system are shown in Table 7. The model uses a custom trained YOLOv12n backbone with about 2.56m parameters, runing at 6.3 GFLOPs. Despite its multitask nature, the system operates in real time on a CPU based architecture with an average inference time of 95–150 ms and a corresponding frame rate of 6–10 FPS. The model size is also relatively small (5.27 MB) and the audio feedback pipeline is non-blocking, both of which are suitable for edge and wearable devices with limited resources.

TABLE 7. Overall System Complexity and Real-Time Performance Analysis

Metric	Value
System Scope	Multi-Feature Assistive System
Supported Features	Object, Footpath, Currency, Face, OCR
Model Backbone	YOLOv12n (Custom Trained)
Inference Layers (Fused)	159
Inference Parameters	∼2.56 M
Computational Cost	6.3 GFLOPs
Input Resolution	384 × 640
Average Inference Time (CPU)	95–150 ms
Frame Rate (CPU)	∼6–10 FPS
Frame Processing Time	∼100–155 ms
Detection-to-Audio Delay	0.6–1.0 ms
Execution Mode	Sequential / Event-driven Pipeline
Audio Feedback	Non-blocking, Event-driven (pygame)
RAM Usage	360–380 MB
Model Size (Disk)	∼5.27 MB
Hardware Platform	CPU-based System

Performance of both known and unknown face recognition module is tabulated in Table 8. Recognition rate is 97.6% with low FRR (3.03%) and no FAR, indicating accurate and no false acceptance of the identification information. It is exactly these qualities that are essential for giving reliable identity-related-instructions to people who are visually impaired.

Multilingual OCR module performance is presented in Tabble 9. The system has an average CER of 0.064 and it can recognize Bangla and English text in real-time. Despite still high WERs because of ornate fonts, lighting variation and motion blur present in video streams, the OCR component offers a readable text output for practical assistive applications.

TABLE 8. Performance Evaluation of Known and Unknown Face Recognition Module

Metric	Value
Total Evaluation Samples	252
Known Face Samples	198
Unknown Face Samples	54
True Positives (TP)	192
False Negatives (FN)	6
False Positives (FP)	0
True Negatives (TN)	54
Recognition Accuracy	97.6%
Recall (Known Faces)	96.97%
False Rejection Rate (FRR)	3.03%
False Acceptance Rate (FAR)	0.00%

TABLE 9. Performance Evaluation of the Real-Time Multilingual OCR Module

Metric	Value
Total OCR Samples (Valid)	123
Skipped Samples (Noise / Incomplete GT)	1
Average Character Error Rate (CER)	0.064
Average Word Error Rate (WER)	0.450
Word Recognition Accuracy (%)	55.01
Mean Confidence Score	0.536
Languages Supported	English + Bangla
OCR Mode	Real-Time Video Stream

We compare the proposed system with the state-of-the-art vision-based assistive aids in Table 10. While existing works concentrate on single or a few functionality, the proposed pipeline combines object detection, footpath safe classification Bangladesh currency identification, face recognition as well as multilingual OCR into an unique real-time pipeline. We note that while the packed feature processing and encoding significantly improve performance, they make our work fundamentally different from previous approaches in terms of integration coverage as well as lightweight design and CPU-based deployment.

Figure 6 shows qualitative results of the proposed YOLO-based multi-feature assistive framework in several indoor and outdoor real-world situations. As depicted in sample outputs, several assistive tasks are simultaneusly realized with same single pipeline : object detection, footpath safety classification, Bangladesh currency recognition and known–unknown face identification and multilingual(Bangla –English) optical character recognition(OCR). In outdoor traffic scenes, high detection confidences sum up with vehicle, pedestrian, tree and road boundaries detected correctly so that the system is able to react well in time on obstacles. indoors, doors, walls, stairs and table tops/chair seats are consistently located for safe navigation in an unknown environment. Its pedestrian walkway detection module reliably classifies the walkways, like Free for Use, Partially Occupied and Not Safe for Use, offering actionable navigation assistance to walking-impaired individuals. In practice, the Bangladesh currency recognition module possesses good generalization ability to recognize several denominations (including similar notes) correctly against various backgrounds and partial occlusion. Face recognition results are displayed in color coded bounding boxes on the basis of whether a known individual or not, promoting socially aware ambient navigation. Further, the OCR module extracts Bangla and English text elements from sign boards, books, posters and public notices; which translates visual textual information into audible feedback.

In general, the sample results confirm that the developed system is robust, flexible and real-time as voice control application. The qualitative results validate that the multi-feature network can work effectively in complex and dynamic environments, which will offer a thorough environmental perception and guiding for visually impaired people.

V. CONCLUSION

In this paper, we introduced an YOLOv12-based multi-feature assistive vision system co-located with a live voice assistant to assist blind people navigating safely and independently. Unlike state-of-the-art assistive systems that address isolated perception tasks, the proposed system integrates environmental object detection, footpath hazard labeling, Bangladesh currency recognition, known and unknown face classification, and multilingual (Bangla–English) OCR using a lightweight single pipeline with modularity. Combining such complementary assistive functions provides a full range of environmental perception and context-aware auditory guidance in everyday indoor and outdoor spaces. Extensive experimental results on various task-specific datasets verify the superiority and reliability of our proposed framework. The system presents state-of-the-art detection and recognition performance, according to currency recognition mAP@0.5 of 0.951, strong assessment for footpath safety, reliable performance to detect navigation-critical object classes, strong recognition accuracy of face with close-to-zero false acceptanc and pragmatic real-time OCR performances on reading Bangla and English text. In addition, computational analysis shows that the framework is efficient in terms of latency and resource usage over a CPU-based platform and can be deployed on edge computing devices or as an app for the wearables.

The qualitative outcomes also demonstrate the robustness of our system in complex and dynamic situations combining multiple assistive tasks natively in equally intuitive voice feedback. Hands-free, low-cognitive-load interaction enables the proposed system to improve situational awareness, user

TABLE 10. Comparison with State-of-the-Art Vision-Based Assistive Systems for Visually Impaired Users

Method	Components Used	Dataset	Functionality / Output	Coverage Area
Hoang et al. [55]	Mobile Kinect, Laptop, Electrode Matrix, Headphone, RF Transmitter	Local dataset	Obstacle detection with audio warning	Indoor
Bai et al. [56]	Depth Camera, Smart Glasses, CPU, Headphone, Ultrasonic Sensor	Not reported	Obstacle recognition with audio output	Indoor
Yang et al. [57]	Depth Camera on Smart Glass, Laptop, Headphone	ADE20K, PASCAL VOC, COCO	Obstacle recognition with directional audio feedback	Indoor / Outdoor
Mancini et al. [58]	RGB Camera, PCB Controller, Vibration Motor	Not reported	Obstacle detection with vibration-based feedback	Outdoor
Bauer et al. [59]	Camera, Smartphone, Smartwatch	PASCAL VOC	Object detection with audio-based guidance	Outdoor
Patil et al. [60]	Sensors, Vibration Motors	No dataset	Obstacle detection with audio alerts	Indoor / Outdoor
Eckert et al. [61]	RGB-D Camera, IMU Sensors	PASCAL VOC	Object detection with audio output	Indoor
Parikh et al. [62]	Smartphone, Server, Headphone	Local dataset (11 objects)	Object detection with audio assistance	Outdoor
AL-Madani et al. [63]	BLE Fingerprinting, Fuzzy Logic	Not reported	Indoor localization assistance	Indoor
Proposed Method (This Work)	**RGB Camera, CPU-based System, Voice Output (Headphone)**	**Custom task-specific datasets (Object, Footpath, Bangladesh Currency)**	**Multi-feature detection including object detection, footpath safety classification, Bangladesh currency recognition, known/unknown face identification, and real-time multilingual OCR with voice feedback**	**Indoor / Outdoor**

confidence and general mobility for visually impaired users.

The current implementation, however, also have some limitations and drawbacks notably, lower OCR accuracy under heavy motion blur and complex lighting conditions, as well as poor generalization to unseen object categories. Future work may include integrating adaptive learning techniques, enlarging diversity of dataset, and introducing more sophisticated multimodal perception modules (e.g., depth estimation and vision–language reasoning). User-centred field studies will also be carried out to further evaluate usability and long-term effectiveness. In summary, this work is a practical and scalable approach towards intelligent real-time assistive navigation system which can change the quality of visually impaired user's life in many folds.

FUTURE WORK

In the next phase, we will further improve its stability and robustness as well as its adaptability to more demanding real-world scenarios. Specifically we will investigate advanced transformer-based OCR models and motion-aware text enhancement approaches to enhance recognition accuracy under heavy motion blur, low illumination or various font styles. We will expand classes and variety of training data used for object detection and footpath analysis modules to enhance the generalization capability on unexpected environments. Additionally, we will explore multimodal perception modules including depth estimation, spatial sound localization and vision–language reasoning to give a more comprehensive context understanding and natural audio guidance. In addition, energy-efficient model optimization and on-device learning technique will be investigated to enhance the performance for low-power wearable platforms. Finally, large-scale user-based field trials with visually impaired users will be carried out to evaluate the long-term usability, reliability and real-world impact of the proposed assistive framework.

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in American English is without an "e" after the "g." Use the singular heading even if you have many acknowledgments. Avoid expressions such as "One of us (S.B.A.) would like to thank . . . ." Instead, write "F. A. Author thanks . . . ." In most cases, sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page, not here.

REFERENCES

[1] W. H. Organization *et al.*, "World report on vision: Executive summary," tech. rep., World Health Organization, 2019.

[2] M. A. Hersh and M. A. Johnson, *Assistive technology for visually impaired and blind people*, vol. 1. Springer, 2008.

[3] S. Sivan and G. Darsan, "Computer vision based assistive technology for blind and visually impaired people," in *Proceedings of the 7th international conference on computing communication and networking technologies*, pp. 1–8, 2016.

[4] Y. Zhao, C. L. Bennett, H. Benko, E. Cutrell, C. Holz, M. R. Morris, and M. Sinclair, "Enabling people with visual impairments to navigate virtual

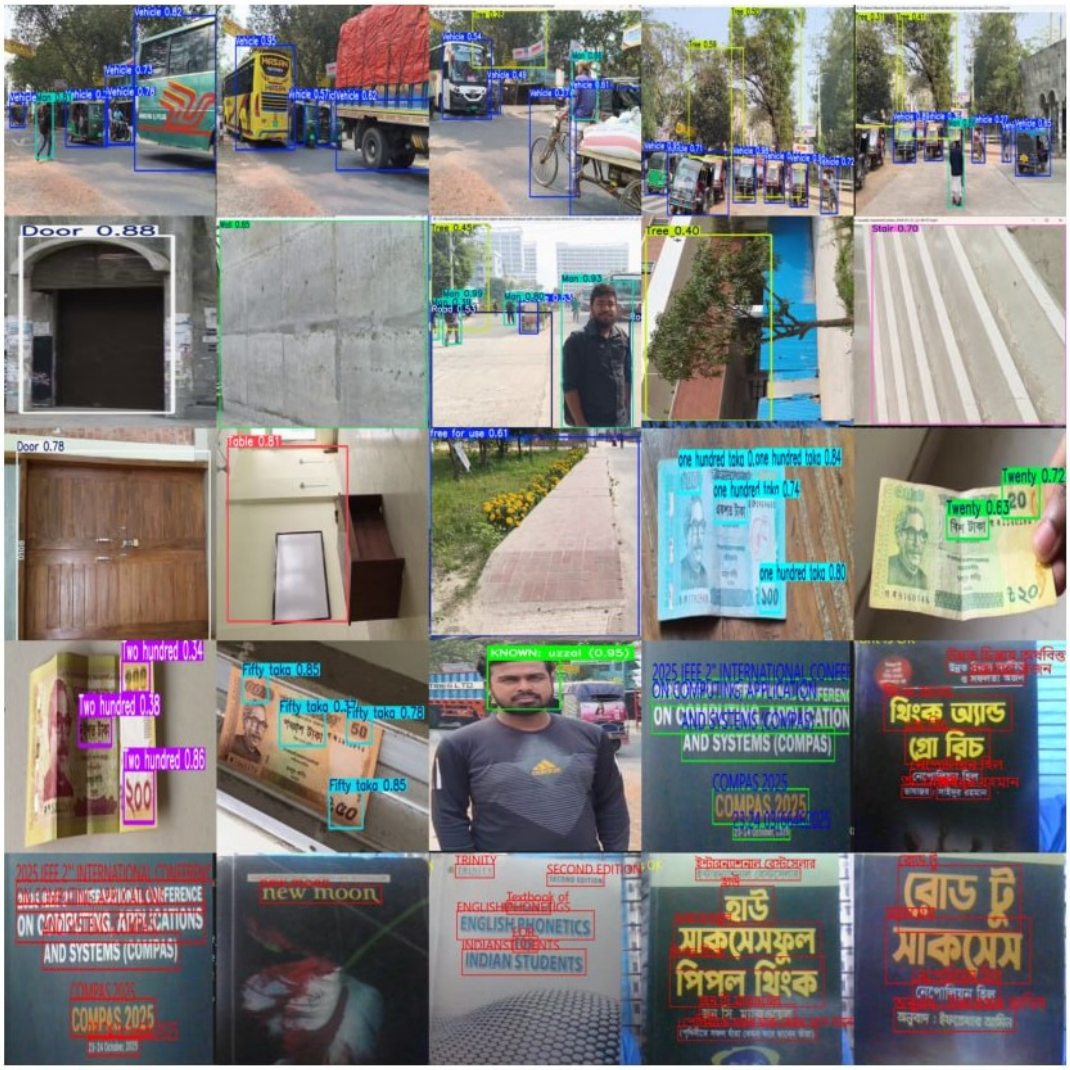

FIGURE 6. **Working flow of the Optical Character Recognition (OCR) system for Bangla and English text**

reality with a haptic and auditory cane simulation," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–14, 2018.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[8] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7464–7475, 2023.

[9] K. Xia, X. Li, H. Liu, M. Zhou, and K. Zhu, "Ibgs: A wearable smart system to assist visually challenged," *IEEE Access*, vol. 10, pp. 77810–77825, 2022.

[10] G. Voutsakelis, I. Dimkaros, N. Tzimos, G. Kokkonis, and S. Kontogiannis, "Development and evaluation of a tool for blind users utilizing ai object detection and haptic feedback," *Machines*, vol. 13, no. 5, p. 398, 2025.

[11] J. Bai, S. Lian, Z. Liu, K. Wang, and D. Liu, "Smart guiding glasses for visually impaired people in indoor environment," *IEEE Transactions on Consumer Electronics*, vol. 63, no. 3, pp. 258–266, 2017.

[12] R. Agarwal, N. Ladha, M. Agarwal, K. K. Majee, A. Das, S. Kumar, S. K. Rai, A. K. Singh, S. Nayak, S. Dey, *et al.*, "Low cost ultrasonic smart glasses for blind," in *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 210–213, IEEE, 2017.

[13] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[14] G. Voutsakelis, I. Dimkaros, N. Tzimos, G. Kokkonis, and S. Kontogiannis, "Development and evaluation of a tool for blind users utilizing ai object detection and haptic feedback," *Machines*, vol. 13, no. 5, p. 398, 2025.

[15] M. T. Islam, M. A. Rashid, A. Mohiuddin, A. Kuwana, and H. Kobayashi, "Design and implementation of smart guided glass for visually impaired people," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 5, p. 5543, 2022.

[16] I. Ouali, M. B. Halima, *et al.*, "Augmented reality for scene text recognition, visualization and reading to assist visually impaired people," *Procedia Computer Science*, vol. 207, pp. 158–167, 2022.

[17] M. M. Valipoor, *Computer Vision Driven Assistive Solution for People with Visual Impairment or Blindness*. PhD thesis, UNIVERSIDAD POLITÉCNICA DE MADRID, 2024.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[19] R. Tapu, B. Mocanu, and T. Zaharia, "Wearable assistive devices for visually impaired: A state of the art survey," *Pattern recognition letters*, vol. 137, pp. 37–52, 2020.

[20] M. Hersh, "Route learning by blind and partially sighted people," *Journal of Blindness Innovation and Research*, vol. 10, no. 2, 2020.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[22] M. A. Rahman and M. S. Sadi, "Iot enabled automated object recognition for the visually impaired," *Computer methods and programs in biomedicine update*, vol. 1, p. 100015, 2021.

[23] M. Hussan, D. Saidulu, P. Anitha, A. Manikandan, and P. Naresh, "Object detection and recognition in real time using deep learning for visually impaired people," *International Journal of Electrical and Electronics Research*, vol. 10, no. 2, pp. 80–86, 2022.

[24] R. B. Islam, S. Akhter, F. Iqbal, M. S. U. Rahman, and R. Khan, "Deep learning based object detection and surrounding environment description for visually impaired people," *Heliyon*, vol. 9, no. 6, 2023.

[25] N. Kumar and A. Jain, "A deep learning based model to assist blind people in their navigation.," *J. Inf. Technol. Educ. Innov. Pract.*, vol. 21, pp. 95–114, 2022.

[26] G. I. Okolo, T. Althobaiti, and N. Ramzan, "Smart assistive navigation system for visually impaired people," *Journal of Disability Research*, vol. 4, no. 1, p. 20240086, 2025.

[27] T. J. Alahmadi, A. U. Rahman, H. K. Alkahtani, and H. Kholidy, "Enhancing object detection for vips using yolov4_resnet101 and text-to-speech conversion model," *Multimodal Technologies and Interaction*, vol. 7, no. 8, p. 77, 2023.

[28] R. Arifando, S. Eto, and C. Wada, "Improved yolov5-based lightweight object detection algorithm for people with visual impairment to detect buses," *Applied Sciences*, vol. 13, no. 9, p. 5802, 2023.

[29] B. Sophia and D. Chitra, "Segmentation based real time anomaly detection and tracking model for pedestrian walkways.," *Intelligent Automation & Soft Computing*, vol. 36, no. 3, 2023.

[30] I. V. Pustokhina, D. A. Pustokhin, T. Vaiyapuri, D. Gupta, S. Kumar, and K. Shankar, "An automated deep learning based anomaly detection in pedestrian walkways for vulnerable road users safety," *Safety science*, vol. 142, p. 105356, 2021.

[31] H. Alsolai, F. N. Al-Wesabi, A. Motwakel, and S. Drar, "Assisting visually impaired people using deep learning-based anomaly detection in pedestrian walkways for intelligent transportation systems on remote sensing images," *Journal of Disability Research*, vol. 2, no. 2, pp. 49–56, 2023.

[32] D. Morra, X. Zhu, C. Liu, K. Fu, F. Duarte, S. Mora, Z. He, and C. Ratti, "Mapping sidewalk accessibility with smartphone imagery and visual ai: a participatory approach," *Philosophical Transactions A*, vol. 382, no. 2285, p. 20240106, 2024.

[33] J. K. Mahendran, D. T. Barry, A. K. Nivedha, and S. M. Bhandarkar, "Computer vision-based assistance system for the visually impaired using mobile edge artificial intelligence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2418–2427, 2021.

[34] K. Veena, B. S. Ullal, P. Gogoi, K. Singh, A. Biswas, and K. Yash, "Smart navigation aid for visually impaired person using a deep learning model," in *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pp. 1049–1053, IEEE, 2023.

[35] R. Lima, L. Barreto, A. Amaral, and S. Paiva, "Visually impaired people positioning assistance system using artificial intelligence," *IEEE Sensors Journal*, vol. 23, no. 7, pp. 7758–7765, 2023.

[36] O. Daescu, H. Huang, and M. Weinzierl, "Deep learning based face recognition system with smart glasses," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 218–226, 2019.

[37] S. Chen, D. Yao, H. Cao, and C. Shen, "A novel approach to wearable image recognition systems to aid visually impaired people," *Applied Sciences*, vol. 9, no. 16, p. 3350, 2019.

[38] W.-J. Chang, L.-B. Chen, C.-Y. Sie, and C.-H. Yang, "An artificial intelligence edge computing-based assistive system for visually impaired pedestrian safety at zebra crossings," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 1, pp. 3–11, 2020.

[39] W.-J. Chang, L.-B. Chen, M.-C. Chen, J.-P. Su, C.-Y. Sie, and C.-H. Yang, "Design and implementation of an intelligent assistive system for visually impaired people for aerial obstacle avoidance and fall detection," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10199–10210, 2020.

[40] U. Kadam, R. Kushwaha, A. Meena, C. Abuzar, Ujjwal, G. Singal, and M. Verma, "Hazardous object detection for visually impaired people using edge device," *SN Computer Science*, vol. 6, no. 1, p. 7, 2024.

[41] A. Bhandari, G. S. Batutis, A. Jain, M. C. Sico, G. Hamilton-Fletcher, C. Feng, T. E. Hudson, J.-R. Rizzo, and K. C. Chan, "Using transfer learning to refine object detection models for blind and low vision users," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–4, IEEE, 2024.

[42] L. C. Méndez-González, "Assistive device for the visually impaired based on computer vision," *Instituto de Ingeniería y Tecnología*, 2023.

[43] A. N. Aniedu, S. C. Nwokoye, C. S. Okafor, K. Anyanwu, and A. N. Isizoh, "Enhanced ai-based navigation system for the visually impaired," *Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 10, no. 1, pp. 16–20, 2025.

[44] R. C. Joshi, S. Yadav, M. K. Dutta, and C. M. Travieso-Gonzalez, "Efficient multi-object detection and smart navigation using artificial intelligence for visually impaired people," *Entropy*, vol. 22, no. 9, p. 941, 2020.

[45] F. E.-Z. El-Taher, A. Taha, J. Courtney, and S. Mckeever, "A systematic review of urban navigation systems for visually impaired people," *Sensors*, vol. 21, no. 9, p. 3103, 2021.

[46] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.

[47] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[48] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[49] W. J. Scheirer, A. Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013.

[50] JaidedAI, "Easyocr: Ready-to-use ocr with deep learning," *GitHub Repository*, 2021.

[51] B. B. Chaudhuri and U. Pal, "A survey on optical character recognition for indic scripts," *Pattern Recognition*, 2020.

[52] R. e. a. Sarkar, "Recent advances in optical character recognition for low-resource languages," *IEEE Access*, 2022.

[53] M. e. a. Li, "Trocr: Transformer-based optical character recognition," *ICCV*, 2021.

[54] R. Atienza, "Scene text recognition with transformer networks," *Pattern Recognition*, 2021.

[55] V. N. Hoang, T. H. Nguyen, T. L. Le, T. H. Tran, T. P. Vuong, and N. Vuillerme, "Obstacle detection and warning system for visually impaired people based on electrode matrix and mobile kinect," *Vietnam Journal of Computer Science*, vol. 4, no. 2, pp. 71–83, 2017.

[56] J. Bai, S. Lian, Z. Liu, K. Wang, and D. Liu, "Smart guiding glasses for visually impaired people in indoor environment," *IEEE Transactions on Consumer Electronics*, vol. 63, no. 3, pp. 258–266, 2017.

[57] K. Yang, L. M. Bergasa, E. Romera, R. Cheng, T. Chen, and K. Wang, "Unifying terrain awareness through real-time semantic segmentation," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1033–1038, IEEE, June 2018.

[58] A. Mancini, E. Frontoni, and P. Zingaretti, "Mechatronic system to help visually impaired users during walking and running," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 2, pp. 649–660, 2018.

[59] Z. Bauer, A. Dominguez, E. Cruz, F. Gomez-Donoso, S. Orts-Escolano, and M. Cazorla, "Enhancing perception for the visually impaired with deep learning techniques and low-cost wearable sensors," *Pattern Recognition Letters*, vol. 137, pp. 27–36, 2020.

[60] K. Patil, Q. Jawadwala, and F. C. Shu, "Design and construction of electronic aid for visually impaired people," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 2, pp. 172–182, 2018.

[61] K. Patil, Q. Jawadwala, and F. C. Shu, "Design and construction of electronic aid for visually impaired people," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 2, pp. 172–182, 2018.

[62] K. Patil, Q. Jawadwala, and F. C. Shu, "Design and construction of electronic aid for visually impaired people," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 2, pp. 172–182, 2018.

[63] B. Al-Madani, F. Orujov, R. Maskeliūnas, R. Damaševičius, and A. Venčkauskas, "Fuzzy logic type-2 based wireless indoor localization system for navigation of visually impaired people in buildings," *Sensors*, vol. 19, no. 9, p. 2114, 2019.

FIRST A. AUTHOR received the B.S. and M.S. degrees in aerospace engineering from the University of Virginia, Charlottesville, in 2001 and the Ph.D. degree in mechanical engineering from Drexel University, Philadelphia, PA, in 2008.

From 2001 to 2004, he was a Research Assistant with the Princeton Plasma Physics Laboratory. Since 2009, he has been an Assistant Professor with the Mechanical Engineering Department, Texas A&M University, College Station. He is the author of three books, more than 150 articles, and more than 70 inventions. His research interests include high-pressure and high-density nonthermal plasma discharge processes and applications, microscale plasma discharges, discharges in liquids, spectroscopic diagnostics, plasma propulsion, and innovation plasma applications. He is an Associate Editor of the journal *Earth, Moon, Planets*, and holds two patents.

Dr. Author was a recipient of the International Association of Geomagnetism and Aeronomy Young Scientist Award for Excellence in 2008, and the IEEE Electromagnetic Compatibility Society Best Symposium Paper Award in 2011.