INTERNATIONAL JOURNAL OF STEM RESEARCH

Science · Technology · Engineering · Mathematics

Manuscript Type: *Original Research*
Published: *[16-07-2026]*
DOI: *https://doi.org/10.19895/ijstemr.2026.1.4*

Visionaid: An Assistive System for Navigation and Environmental Awareness in Visually Impaired Users

Thammarat Vongsawat[1*], Kachornattapon Khumyaito[1], Parin Pongaen[1]

[1] *English Program Yothinburana School, Bangkok, Thailand*

ABSTRACT

Blind and visually impaired individuals face persistent barriers to independent travel, financial transactions, and social recognition, and existing assistive aids typically address these needs in isolation while often depending on cloud connectivity. This study designs and evaluates a low-cost, wearable, multi-modal assistive device that consolidates obstacle avoidance, guide-path following, banknote identification, and acquaintance recognition, with all artificial-intelligence inference performed on the device itself. The prototype was built around a Raspberry Pi single-board computer and an OAK-D Lite stereo-depth smart camera, coordinated by the Robot Operating System (ROS 2), and provides three voice-selectable modes: a walking mode combining stereo-depth obstacle detection with colour-based tactile-paving guidance; a banknote mode using an on-device object-detection network for five Thai denominations; and a face mode identifying a pre-enrolled set of individuals. Interaction is fully non-visual, relying on Thai and English speech commands with spoken text-to-speech feedback, and the recognition models were evaluated on labelled image sets. The banknote model achieved a macro mAP@0.5 of 0.966 across five denominations (macro precision 0.972, recall 0.879), with the ฿100 class showing the highest confusion rate; the face module achieved 99.9% accuracy over three enrolled individuals; and obstacle detection reliably flagged objects within 1.5 metres across left, centre, and right zones at 30 frames per second. These results demonstrate the feasibility of an integrated, on-device, voice-driven wearable that delivers multiple independence-supporting functions on affordable embedded hardware.

 | DOI: *https://doi.org/10.19895/ijstemr.2026.1.4*

Keywords: assistive technology; visually impaired persons; wearable electronic devices; edge computing; stereo vision; speech interfaces

1. INTRODUCTION

1.1 Background

Vision is the primary channel through which people perceive and navigate the physical and social world, and its loss imposes daily constraints on independence. Globally, at least 2.2 billion people live with a vision impairment [C1], and visual disability is among the largest registered disability groups in Thailand [C2]. For blind and visually impaired (BVI) individuals, three everyday activities remain disproportionately difficult and consequential for autonomy: travelling safely on foot, handling cash, and recognising the people nearby [C3]. Existing aids address these needs only partially. The white cane and guide dog detect only near, ground-level hazards [C4]; ultrasonic electronic travel aids extend range but report a single distance without spatial context [C5]; identifying banknotes by touch is error-prone [C6]; and conventional aids do not support social recognition at all [C7]. Camera-based systems can, in principle, perceive scene geometry, read currency, and identify people [C8], yet most are single-function [C9], depend on cloud inference with its connectivity, latency, privacy, and cost burdens [C10], and are seldom designed for resource-constrained, non-Western contexts such as Thailand.

1.2 Objectives

The objectives of this study are:

1. To design and implement a wearable, on-device, multi-modal assistive system integrating (a) stereo-depth obstacle detection with tactile-paving guidance, (b) Thai banknote identification, and (c) recognition of a pre-enrolled set of acquaintances, under a unified voice-driven interface.
2. To characterise the performance of the on-device recognition models on labelled image data and the behaviour of the depth-based obstacle module.
3. To assess the feasibility of consolidating these functions on affordable embedded hardware operating without cloud connectivity, and to identify the limitations to be resolved before field deployment.

1.3 Significance

This study contributes a single, low-cost wearable that consolidates multiple independence-supporting functions while performing all inference on the device itself, addressing three gaps in existing aids. Its design embodies three principles: *edge-first computing*, which removes network dependence and preserves the privacy of continuously observed surroundings [C11]; *class-agnostic obstacle sensing*, which uses stereo depth directly so that any physical obstacle—wall, pole, glass, or box—is detected whether or not it has been seen before [C12]; and *non-visual, multimodal interaction*, in which the entire interface is auditory, pairing speech-command control with spoken feedback in the user's language. By targeting Thai currency, tactile-paving conventions, and language, the work also demonstrates assistive technology adapted to a non-Western, resource-constrained setting.

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

1.4 Scope of Study

This paper presents and evaluates a working prototype with preliminary results. The scope covers the design of the integrated hardware-software system and a quantitative technical evaluation of its perception modules on labelled image datasets and controlled bench/walkthrough conditions. The recognition component is limited to five Thai banknote denominations and a closed, pre-enrolled set of three individuals; the face module performs fixed-roster identification rather than open-set recognition. A controlled trial with BVI end-users is outside the present scope and is identified as future work. The remainder of this paper is organised as follows: Section 2 reviews related work; Section 3 details the methodology; Section 4 reports results; Section 5 discusses implications and limitations; and Section 6 concludes.

2. LITERATURE REVIEW

Wearable assistive technologies for visually impaired users have been systematically classified by sensor modality – camera, LiDAR, and ultrasonic – with camera-based systems augmented by depth sensing identified as offering the most comprehensive scene understanding within a miniaturised wearable form factor [R7]. A persistent limitation across surveyed prototypes, however, is their single-function scope: most systems address either navigation or object recognition, but seldom both within one integrated device [R7].

The perception capabilities of such systems have been transformed by the evolution of deep convolutional neural networks and, more recently, Vision Transformers. [R1] establishes that hierarchical backbone architectures pre-trained on large-scale datasets produce transferable feature representations suitable for fine-tuning on specialised downstream tasks. The current trend toward computationally efficient backbone designs – reducing multiply-accumulate operation counts without sacrificing benchmark accuracy – has made real-time inference feasible on edge hardware lacking discrete GPUs [R1], a prerequisite for any wearable deployment.

Within the recognition domain, currency identification and human detection represent two high-value tasks for visually impaired users. Bandara [R2] demonstrated high-precision banknote recognition using EfficientDet-lite2 with a weighted bidirectional feature pyramid network, achieving a mean average precision of 0.9877 across six Sri Lankan denominations and delivering results via on-device text-to-speech, yet the system remains restricted to smartphone deployment and a single country's note series. At an earlier stage of the field, Rajesh et al. [R3] combined Tesseract OCR with Haar cascade face detection on a Raspberry Pi 3, demonstrating that multi-modal recognition is achievable on low-cost embedded hardware. Both methods, however, suffer from well-documented degradation under uncontrolled outdoor illumination and non-frontal viewing angles, limiting practical deployment to controlled indoor environments.

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

For navigation, stereo-vision obstacle detection has evolved from FPGA-based disparity processors [R5] – which highlighted the fundamental resolution-latency trade-off in dense depth map computation – toward sensor-fusion approaches that combine disparity with appearance-based segmentation. Dzhans and Tolstaya [R4] addressed the failure case of low-contrast ground-level obstacles by fusing SLIC superpixel segmentation with stereo depth data, reliably distinguishing three-dimensional objects from flat patterns such as shadows and road markings that mislead appearance-only detectors. Tactile paving detection, directly relevant to guided pedestrian navigation across Asia, has been addressed by Li et al. [R6] using a lightweight architecture combining an Improved G-GhostNet backbone with Coordinate Attention and a redesigned atrous spatial pyramid pooling module, achieving 94% mean intersection over union at 59.2 fps on an embedded NPU – a strong accuracy-speed benchmark for this class of problem.

Despite these advances, no existing system unifies currency recognition, obstacle avoidance, tactile-paving guidance, face identification, and voice interaction within a single wearable prototype. Furthermore, no prior work addresses the Thai operational context specifically – Thai Baht denominations, Thai-standard guiding tiles, or a Thai-language audio interface. The present study addresses this gap.

3. RESEARCH METHODOLOGY

3.1 Research Design

This study adopts a design-and-development (engineering) research design combined with a quantitative technical evaluation, rather than an experimental or hypothesis-testing design. This choice is dictated by the research objectives: the central aim is to design, build, and demonstrate the feasibility of an integrated, on-device, multi-modal assistive wearable, and then to characterise the performance of its constituent perception modules under controlled conditions. The work therefore proceeds in two phases. The first is a constructive phase, in which system requirements are derived from the daily-living needs of blind and visually impaired (BVI) users and translated into a modular hardware-software architecture. The second is an evaluative phase, in which the recognition models (banknote and face) are assessed quantitatively on labelled image data, and the depth-based obstacle module and tactile-paving guidance module are assessed descriptively under controlled bench and walkthrough conditions. Because this paper reports a working prototype with preliminary results, the design deliberately excludes a controlled trial with BVI end-users at this stage; such a field study is identified as future work (Section 5).

3.2 Participants / Samples

This phase of the research involved no human BVI participants; consequently, no inferential power analysis or participant sampling was required. The units of analysis are instead three categories of the evaluation sample:

1. **Banknote image dataset:** Images of the five Thai Baht denominations were captured using the OAK-D Lite colour camera (CAM_A) under controlled indoor lighting, yielding 76 raw images across five classes (฿20: 11, ฿50: 13, ฿100: 11, ฿500: 22, ฿1,000: 19). Images were annotated in YOLO bounding-box format using CIRA Core offline annotation software. Offline augmentation (rotation, contrast adjustment, Gaussian noise, Gaussian blur) at ×22 per c

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

lass expanded the dataset to 1,672 images, split into 1,505 training / 167 validation images (~90/10). [C16]

2. **Face image dataset:** A purpose-built dataset of 213 raw images across three enrolled individuals – book (70 images), pearn (81 images), and tonnam (62 images). These were constructed using the OAK-D Lite colour camera under varied indoor lighting, distances (approximately 0.5-2 m), and head angles. Offline augmentation ×5 per class yielded 935 training / 103 validation images (~90/10 split). The three individuals are members of the research team who provided written informed consent.
3. **Obstacle and tactile-path test scenarios:** A set of controlled scenarios used to characterise the walking mode. Formal controlled obstacle and tactile-paving trials were not conducted in the present study; the walking-mode modules were characterised through informal observation during development, as described in §4.3 and §4.4. Quantitative controlled trials are identified as a primary direction for future work.

The dataset sizes follow common practice for transfer-learning-based object detectors on narrow, closed-set tasks, where a few hundred well-labelled images per task are typically sufficient for convergence [C13].

3.3 Instruments and Materials

3.3.1 Hardware

The hardware platform of the proposed wearable assistive system comprises four principal subsystems: a stereo-vision camera unit with on-chip neural processing, a single-board host computer for ROS 2 orchestration, an audio peripheral pair for voice input and spoken output, and a portable power supply. All components are mounted on a vest-shoulder rig so that the camera maintains an unobstructed forward field of view above pelvis level to ensure reliable obstacle detection. The assembled prototype is shown in Figure 1, and the electrical connections between subsystems are illustrated in Figure 2.

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

Fig. 1 – photo of assembled prototype, worn by a research member

The electrical connections and data flow between all subsystems are summarised in Figure 2.

Fig. 2 – hardware architecture/block diagram (that PowerPoint diagram)

The primary sensing and inference unit is the OAK-D Lite (Luxonis, Inc.), a compact spatial AI camera with a front face of 91 × 28 mm and a body depth of 17.5 mm (Figure 3). Three imaging sensors are integrated within the single housing: one central RGB colour camera (Sony IMX214, 13 MP, 81° diagonal field of view) for appearance-based recognition tasks, and two monochrome stereo cameras (OmniVision OV7251, 640 × 480 resolution, 89.5° diagonal field of view) separated by a 75 mm baseline for stereo depth computation. The OAK-D Lite incorporates an Intel Movidi

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

us Myriad X (RVC2) vision processing unit (VPU), which provides 4 TOPS of dedicated neural inference throughput and executes all convolutional network operations entirely on-chip at 30 frames per second, offloading the host processor from inference workloads entirely. The device connects to the host computer via a USB 3.1 Gen 1 Type-C interface, through which both power delivery and bidirectional data transfer are handled over a single cable. Because the device is mounted in an inverted orientation on the wearable rig, the image orientation is corrected in firmware by setting ROTATE_180_DEG via the DepthAI pipeline API, so all downstream detection coordinates are reported in the correct upright frame. Fixed focus is applied to the colour camera at a lens position of 15 (on a 0-255 scale, where 0 = infinity and 255 = macro), calibrated empirically to maintain sharpness at typical walking distances of 0.5-3 m while preventing the autofocus mechanism from hunting during motion.

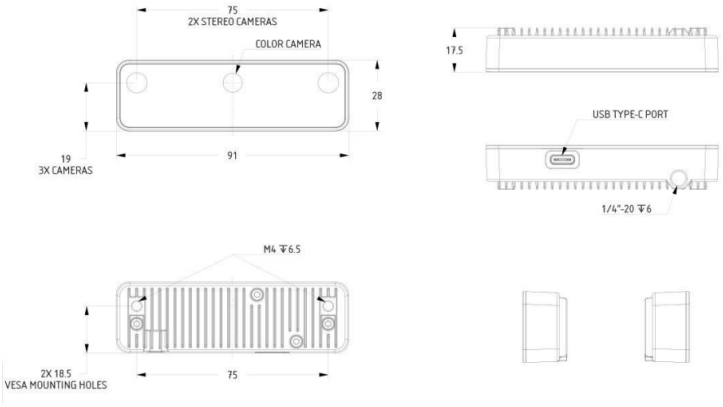


Fig. 3 – OAK-D Lite mechanical drawing. ***Cited:*** *link*

The host processing unit is a Raspberry Pi 5 running Ubuntu 24.04 and the ROS 2 Jazzy middleware. The Raspberry Pi executes the four ROS 2 nodes responsible for orchestration, decision logic, audio output, and web display (§3.3.2), receiving pre-decoded detection results from the OAK-D Lite over USB rather than performing any neural inference locally. This deliberate separation of responsibilities – neural inference on the VPU, logic and inter-process communication on the host CPU – maintains host processor utilisation within a manageable range during full five-function operation and leaves sufficient headroom for concurrent node execution. The host is also responsible for running the live JPEG preview stream, compressed at quality level 50 and downscaled to a maximum width of 480 pixels, which is pushed to the web dashboard at port 8080 for caregiver or researcher monitoring.

Voice input is captured via an Essager USB/Type-C external sound card connected to a 3.5 mm headset microphone operating at a sample rate of 16,000 Hz, compatible with both the PyAudio and ALSA (arecord) capture backends supported by the speech_logger_node. The recogniser applies a 1-second ambient noise calibration window before each listening cycle and enforces an RMS energy threshold of 120 to reject silent audio frames, reducing spurious API calls. Speech transcription is performed by the Google Speech Recognition API, configured for Thai (th-TH) as the primary language with a fall-through to English (en-US) selectable at launch. Audio output is delivered through 3.5 mm earphones via the same Essager USB/Type-C sound card, driven by Microsoft Edge TTS (ed

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

ge-playback) using the th-TH-PremwadeeNeural neural voice for Thai and en-US-EmmaNeural for English. To prevent overlapping announcements, a first-in, first-out speech queue is maintained in a dedicated thread; any queued utterance beyond the first is discarded when a new, higher-priority result arrives, ensuring the user always hears the most current detection without audio backlog.

The entire system is powered by a 3,300 mAh, 12.6 V LiPo battery (3C2P configuration) connected to the Raspberry Pi via USB-C Power Delivery, providing an estimated runtime of approximately 3 hours under continuous 30 fps inference. The OAK-D Lite draws power directly from the Raspberry Pi's USB 3.0 port, which supplies up to 900 mA at 5 V – within the camera's rated power envelope. The complete wearable has a total component cost of approximately 15,000 THB, itemised in Table 3.

3.3.2 Software.

The software architecture of the proposed system is built on the Robot Operating System 2 (ROS 2) middleware framework, which provides a standardised publish-subscribe communication model, lifecycle management for sensor nodes, and parameter-driven configuration without code changes. ROS 2 was selected over ROS 1 for its native support for real-time communication, improved security, and active long-term support releases that target Linux on ARM single-board computers. The entire system is implemented in Python 3 and comprises four ROS 2 nodes that exchange data exclusively through typed message topics, achieving a fully decoupled architecture in which any node can be restarted or replaced independently.

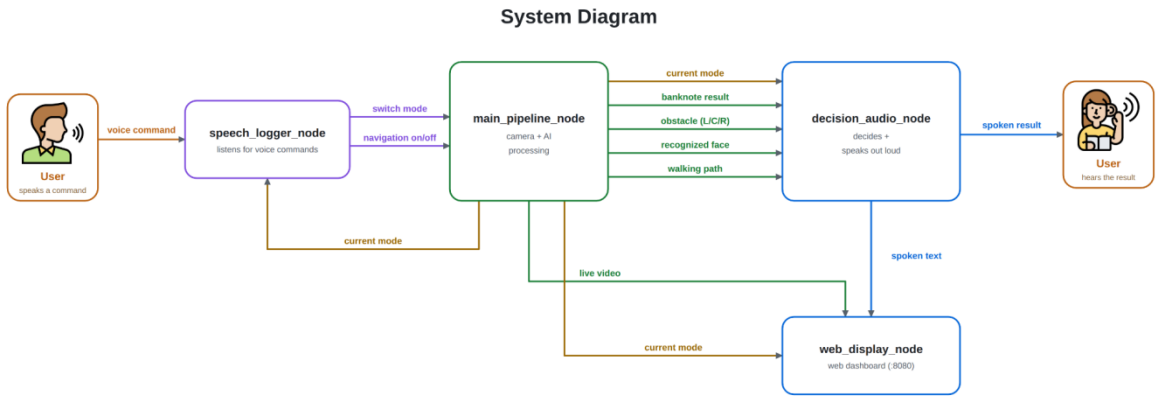


Fig. 4a. ROS 2 node-topic communication graph of the proposed system. Arrows indicate published topics; colours denote the publishing node.

The four nodes and their responsibilities are as follows. The main_pipeline_node owns the OAK-D Lite camera pipeline and serves as the sole interface to the DepthAI 3.x API. Because the OAK-D Lite supports only one active pipeline at a time, this node is responsible for tearing down the current pipeline and building a new one whenever the operating mode changes; a mode transition takes approximately 1-2 seconds during which the previous pipeline is closed, and the USB device is reopened. The speech_logger_node captures audio from the microphone, transcribes voice commands using the Google Speech Recognition API at 16,000 Hz, and publishes mode-switch or navigation-toggle commands onto the appropriate topics. The decision_audio_node receives all detection results from the main_pipeline_node, applies priority-based filtering and temporal voting logic,

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

and drives the text-to-speech (TTS) engine to produce spoken output for the user. The web_display_node subscribes to the live JPEG preview and the latest spoken announcement, and serves a browser-accessible dashboard on port 8080 for monitoring purposes.

Topic-based communication between nodes is carried over the following ROS 2 topics. The speech_logger_node publishes mode commands to /oak/mode_cmd and navigation on/off commands to /oak/nav_cmd, both using the custom oak_interfaces/msg/Mode and std_msgs/Bool message types, respectively. The main_pipeline_node broadcasts the currently active mode on /oak/current_mode every second as a heartbeat, ensuring all subscribers remain synchronised even across reconnections. Detection results are published on four separate topics – /money/detections, /navigation/detections, /face/detections, and /tactile/path – all using the custom oak_interfaces/msg/Detection2DArray message type, which encodes class name, confidence score, and normalised bounding-box coordinates. The live camera preview is published on /oak/preview/compressed as a sensor_msgs/CompressedImage JPEG, and the latest spoken phrase is mirrored to /oak/announcement as a std_msgs/String for the web dashboard. A /oak/language topic allows the system language (Thai or English) to be switched at runtime without restarting any node.

Operating modes are defined by a three-state enumeration WALK (mode 0), MONEY (mode 1), and FACE (mode 2) – declared in oak_interfaces/msg/Mode. The system starts in WALK mode by default. When main_pipeline_node receives a mode command, it calls _cleanup_pipeline() to close the current DepthAI device connection and clear all output queues, then immediately builds and starts the pipeline for the new mode. In WALK mode, the pipeline activates the stereo cameras (left: CAM_B, right: CAM_C) at 640 × 400 pixels with the HIGH_DENSITY depth preset, aligned to the colour camera frame (CAM_A), and additionally streams 640 × 360 BGR frames from CAM_A for the tactile detection module. In MONEY and FACE modes, the pipeline switches to a single-camera DetectionNetwork configuration that loads the respective YOLO11-nano NN Archive and runs on-chip inference at 30 fps; both modes use the identical pipeline structure with only the archive file and per-class confidence thresholds differing.

Inference polling is performed by a 50 ms ROS 2 timer (20 Hz) within the main_pipeline_node. On each tick, the node drains its output queue fully and retains only the most recently arrived packet, discarding any stale frames that accumulated between ticks. This drain-to-latest strategy prevents detection results from building up a lag during transient processing slowdowns. For banknote and face detection, a confirmed result requires a majority vote across a sliding window of five consecutive frames (money_confirm_count: 5) before the detection is published, reducing the impact of single-frame false positives.

Speech management in decision_audio_node is handled by a dedicated background thread that enforces four rules simultaneously: (i) *coalescing* – only the most recently enqueued utterance per logical key (e.g., "dir" for direction, "alert" for obstacles) is retained, so a rapid burst of identical events never produces a speech backlog; (ii) *prioritisation* – a higher-priority utterance kills the currently playing audio mid-sentence via subprocess.terminate() before starting its own; (iii) *deduplication* – the same phrase on the same key is not repeated within a configurable minimum interval; and (iv) *expiry* – utterances older than their time-to-live (TTL) are discarded unspoken, so a navigation instruction for a turn already passed is never announced late. Priority levels are assigned as: direction guidance (10) < status information (20) < mode changes (30) < recognition results (40) < obstacle alerts (50) < battery undervoltage warning (60). Audio is synthesised using Microsoft Edge TTS (edge-tts) with the th-TH-PremwadeeNeural neural voice f

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

or Thai and en-US-EmmaNeural for English. To minimise per-utterance latency, synthesised phrases are pre-cached at boot time as MP3 files identified by MD5 hash and played by mpv or ffplay locally, requiring a network connection only on the first synthesis of each unique phrase. The system additionally monitors the Raspberry Pi supply voltage via vcgencmd get_throttled every 10 seconds and issues a spoken low-battery warning (repeating at most once per 60 seconds) when the under-voltage bit is set, providing the user with a hardware safety alert without requiring any visual display.

3.3.3 Models.

The proposed system employs three distinct perception models, each dedicated to one functional task. Two of the three are neural-network detectors based on the YOLO11-nano architecture; the third is a classical computer vision pipeline that requires no learned weights. All models operate exclusively in their designated operating mode and are never active simultaneously, as the OAK-D Lite supports only one DepthAI pipeline at a time.

Banknote detection model.

The banknote dataset was constructed by photographing physical Thai Baht notes using the OAK-D Lite colour camera (CAM_A) under controlled indoor lighting, with varied presentation angles and distances to simulate real-world handling. The 76 raw images (฿20: 11, ฿50: 13, ฿100: 11, ฿500: 22, ฿1,000: 19) were annotated in YOLO format using CIRA Core offline annotation software and split 90/10 into training and validation sets. Offline augmentation – rotation, contrast adjustment, Gaussian noise, and Gaussian blur – was applied at ×22 per class via CIRA Core, yielding a final dataset of 1,505 training / 167 validation images.

Face recognition model.

The face recognition dataset comprises 213 raw images across 3 identity classes (book: 70, pearn: 81, tonnam: 62), collected by capturing frames from the OAK-D Lite colour camera under varied indoor lighting, head angles, and distances, ensuring that the training distribution matched the deployment sensor. The dataset was split 90/10 and augmented at ×5 per class (rotation, contrast, noise, blur) via CIRA Core, yielding 935 training / 103 validation images.

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

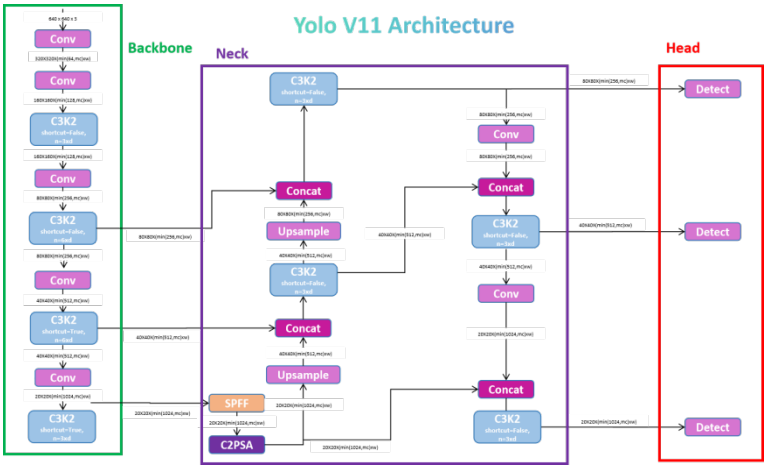


Fig. 4b – YOLO11-nano architecture diagram

Tactile-paving detection model.

The tactile-paving guidance module does not use a neural network. Instead, it applies a classical colour-segmentation pipeline implemented in OpenCV, chosen for its deterministic latency and zero inference overhead on the host CPU. The pipeline processes 640 × 360 BGR frames from the OAK-D Lite's colour camera. Each frame is first converted to the HSV colour space, and a binary mask is produced by thresholding within the yellow hue range $H \in [5, 20]$, saturation $S \geq 25$, and brightness $V \geq 85$ (OpenCV 8-bit scale, where $H \in [0, 179]$). The mask is refined by two sequential morphological operations: an opening pass with a 5 × 5 elliptical structuring element removes isolated noise pixels, and a closing pass with a 7 × 17 elliptical kernel (taller than wide) reconnects broken vertical tile segments; a final 5 × 5 median blur suppresses per-frame speckle. Analysis is confined to the lower 30% of the frame height and the central 30% of the frame width, eliminating sky, distant background, and roadside objects from consideration. The detailed heading-classification procedure built on top of this mask is described in § 3.4.4.

3.3.4 Validity and reliability.

Validity in the context of the proposed wearable system refers to the degree to which each module's output correctly represents the physical state it is intended to measure. For the banknote detection module, validity is defined as the correspondence between the spoken denomination and the denomination physically presented to the camera. For the obstacle detection module, validity is the correspondence between an issued alert and the actual presence of a physical object within 1,500 mm of the device along the reported left, centre, or right zone. For the tactile-paving module, validity is the correspondence between the reported heading (straight, bear, or sharp left/right) and the actual geometric direction of the guiding tile path visible in the camera frame. For the face recognition module, validity is the correspondence between the announced identity and the person physically present.

Several design decisions in the system address construct validity directly. The stereo depth measurement underlying obstacle detection is geometrically grounded: disparity values are converted to millimetre distances by the OAK-D Lite's on-chip stereo engine using the factory-cali

brated camera baseline of 75 mm, making the depth value a physically interpretable quantity rather than an abstract score. The tactile heading is derived from the normalised lateral offset of the fitted path centreline relative to the frame's fixed reference point (frame centre-bottom), which corresponds geometrically to the user's standing position; this offset has a direct physical interpretation as the angular deviation from the path. Confidence thresholds for the neural-network detectors are set conservatively at 0.20 for most banknote classes and 0.70 for face recognition, so that only detections with sufficient evidence are published, trading recall for precision to avoid communicating incorrect information to a visually impaired user.

Reliability refers to the consistency of the system's outputs under repeated exposure to equivalent conditions. Three temporal-smoothing mechanisms are implemented to improve frame-to-frame reliability. First, the banknote detector requires a majority vote across a sliding window of five consecutive frames before a denomination is announced, so a single misclassified frame cannot trigger a false output. Second, the tactile heading is determined by majority vote over all frames received within the preceding one second (minimum five samples required), with an additional 2.5-second dwell constraint that prevents the announced direction from oscillating between adjacent categories when the path image is ambiguous. Third, obstacle zone classification uses the 20th percentile depth value of near pixels within each zone rather than the minimum, making the measurement robust to the isolated erroneous disparity values (flying pixels) that stereo matching routinely produces at object edges. Additionally, when an obstacle is detected, tactile guidance is suppressed for two seconds to prevent the lower-body region of a nearby person from being misread as a guiding tile, a known failure mode of colour-based segmentation in crowded scenes.

Quantitative evidence of validity and reliability, including per-class detection accuracy, confusion matrices, and false positive rates measured during controlled evaluation trials, is presented in § 4.

3.3.5 Novel instrument.

The instrument developed in this study is a unified, voice-controlled wearable assistive device that integrates five perception functions: obstacle avoidance, banknote recognition, tactile-paving guidance, face identification, and voice interaction within a single hardware unit operated through spoken commands in both Thai and English. As established in §2, no prior work has combined these functions in one prototype, and no existing assistive device has been validated against the Thai operational context specifically. The instrument therefore represents a novel contribution both in its functional scope and in its localisation.

The primary technical novelty of the instrument lies in its execution architecture. All neural inference executes on the Intel Movidius Myriad X VPU embedded within the OAK-D Lite camera unit, connected to the host single-board computer over a single USB 3.1 cable. This design eliminates the need for a discrete GPU, cloud connectivity, or a laptop-class processor, reducing the host computer to an orchestration role only. The consequence is a self-contained device whose battery runtime is not constrained by GPU power draw and whose inference latency is not subject to network variability. The ROS 2 publish-subscribe architecture further enables the five functional modules to operate as independent, replaceable nodes that share a common message bus, a design that allows any individual module to be updated, swapped, or disabled without affecting the remaining f

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

unctions. The mode-switching mechanism, driven entirely by the user's spoken commands, allows th e same camera and VPU hardware to serve three distinct inference pipelines sequentially, making t he single OAK-D Lite equivalent in functional coverage to what would otherwise require three sep arate specialised sensors.

A secondary novelty is the instrument's Thai-language integration throughout the full per ception-to-output chain. The banknote detector is trained exclusively on Thai Baht denominations (฿20, ฿50, ฿100, ฿500, ฿1000). The tactile-paving detector is calibrated to the yellow hue of Th ai-standard guiding tiles. The speech recognition engine is configured for Thai (th-TH) as the p rimary language, and audio output uses a Thai neural TTS voice (th-TH-PremwadeeNeural). This end -to-end Thai localisation distinguishes the instrument from all reviewed prior works, which addr ess other national contexts or provide no language-specific adaptation.

3.4 Data Collection Procedures

3.4.1 Model training

Two custom object detection models were trained for the proposed system: a banknote denom ination detector and a face recognition detector. Both adopt the YOLO11-nano architecture [C13], selected for its balance between parameter count and detection accuracy at the input resolution r equired for RVC2 on-chip inference. Training followed an identical pipeline for both models, dif fering only in dataset content, class definitions, and per-class confidence thresholds applied a t inference time.

Dataset construction.

The banknote dataset comprises 76 raw images across five Thai Baht denomination classes (฿20: 11, ฿50: 13, ฿100: 11, ฿500: 22, ฿1,000: 19), collected by photographing physical notes und er varied illumination conditions, backgrounds, and orientations to simulate real-world handling . Images were captured using the OAK-D Lite colour camera (CAM_A) and annotated in YOLO format u sing CIRA Core offline annotation software.

The face recognition dataset comprises 213 raw images across 3 identity classes, collect ed by capturing video frames of 3 individuals (book, pearn, and tonnam) under varied lighting, d istances, and head angles using the OAK-D Lite colour camera directly, ensuring that the trainin g distribution matched the deployment sensor.

Augmentation.

Both datasets underwent a two-stage augmentation pipeline. In the first stage, offline au gmentation was applied using CIRA Core prior to the train/validation split: rotation, contrast a djustment, Gaussian noise, and Gaussian blur at ×22 per class for the banknote dataset and ×5 p er class for the face dataset, expanding the raw collections to 1,672 and 1,038 images respectiv ely.

In the second stage, Ultralytics online augmentation was applied stochastically during training. For the banknote model, several parameters were customised beyond their defaults to reflect the a ppearance variation expected in real-world note handling: geometric jitter included rotation up t o ±15° , shear of 2.0, perspective distortion of 0.0005, scale variation in the range 0.1×-1.9

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

×, and translation of 0.15; vertical flip was disabled (flipud = 0) because Thai Baht notes carry orientation-dependent visual features; colour jitter used hsv_h = 0.02, hsv_s = 0.8, and hsv_v = 0.5; mixup blending (0.15) and copy-paste augmentation (0.3) were enabled to increase inter-class diversity; and mosaic was applied at probability 1.0 and disabled for the final 15 epochs to allow the model to fine-tune on clean single-image samples. For the face recognition model, Ultralytics default online augmentation parameters were retained (scale = 0.5, hsv_h = 0.015, hsv_s = 0.7, hsv_v = 0.4, mosaic = 1.0, randaugment), with no geometric distortions beyond horizontal flip, as face recognition requires stable spatial structure. The full augmentation parameter sets for both models are listed in Table 4.

Training procedure.

Both models were trained using the Ultralytics YOLO11 framework [C13] with the AdamW optimiser and a cosine learning rate schedule over 150 epochs with a batch size of 64. The banknote model used an initial learning rate of 0.00111 and an input resolution of 416 × 416 pixels. The face recognition model used an initial learning rate of 0.001429 and was trained at 640 × 640 pixels; the checkpoint was subsequently converted to a 416 × 416 input NN Archive at deployment (§3.4.1), meaning the two models share the same on-device input resolution despite differing training resolutions. Training was conducted on a local workstation with an NVIDIA RTX 5060 Ti GPU (16 GB VRAM). The best-performing checkpoint, selected by the highest validation mAP@0.5, was saved as best.pt.

Table 1: Training hyperparameters for the banknote denomination and face recognition models.

parameter	ธนบัตร	หน้าเพื่อน
Optimizer	Auto (AdamW)	auto(AdamW)
Learning Rate	**0.00111**	**0.001429**
Epochs	150	150
Batch Size	64	64
momentum	0.9	0.9
imgsz	416	640

3.4.2 Recognition evaluation.

Both the banknote and face recognition models were evaluated using an offline, on-device protocol that mirrors the conditions of live deployment as closely as possible. Rather than exporting weights to an ONNX or CPU runtime, the evaluation pipeline feeds labelled still images directly into the deployed NN Archive (.rvc2.tar.xz) running on the physical OAK-D Lite via a host input queue, so the scores measured reflect the actual quantised model executing on the Myriad X VPU – not an approximation. This design choice ensures that any quantisation error introduced during the tools.luxonis.com conversion is captured in the reported metrics.

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

The evaluation dataset is organised as a flat directory of class sub-folders, where each sub-folder name is treated as the ground-truth label. For the banknote model, sub-folders correspond to the five denomination classes (฿20, ฿50, ฿100, ฿500, ฿1000); for the face model, sub-folders correspond to the 3 registered identity classes. Each image is pre-processed identically to the live pipeline: a centre-crop to the largest inscribed square is applied first, followed by bilinear downscaling to 416 × 416 pixels, replicating the resizeMode=CROP behaviour of the DepthAI DetectionNetwork node. The same per-class confidence thresholds used at runtime (§3.3.3) are applied during evaluation, so an image for which no detection clears its class threshold is recorded as a none (no-detection) prediction rather than a forced class assignment.

The task is treated as closed-set classification: the highest-confidence detection that clears its per-class threshold is taken as the model's prediction for that image, and this prediction is compared against the folder-level ground-truth label. From the resulting prediction list, the following metrics are computed: overall accuracy, per-class precision, recall, F1-score, and support. A confusion matrix is constructed with rows indexed by true label and columns by predicted label, and is saved in both raw-count and row-normalised forms. All output per-image predictions, per-class metrics, and confusion matrix images are written to a structured output directory for inclusion in §4.

The reported numbers constitute validation-set performance, evaluated on 167 images for the banknote model and 103 images for the face model. The complete per-class results and confusion matrices are presented in §4.

3.4.3 Obstacle-detection procedure.

The obstacle-detection module operates exclusively in WALK mode and relies on stereo depth data alone; no neural network is involved. This design choice means the module detects any physical object that displaces space within range, regardless of its appearance: walls, glass panels, bollards, parked vehicles, and people are all detected identically, which is precisely the property required for safe pedestrian navigation.

Stereo depth pipeline.

When WALK mode is active, the main_pipeline_node constructs a DepthAI StereoDepth node from the two monochrome cameras (CAM_B left, CAM_C right), each streaming 640 × 400 pixel NV12 frames at 30 fps. The StereoDepth node is configured with the HIGH_DENSITY preset, which maximises disparity map coverage at the cost of some spatial resolution, and depth output is spatially aligned to the colour camera frame (CAM_A) so that pixel coordinates in the depth map correspond directly to the same scene positions visible in the RGB preview. The resulting depth frame encodes per-pixel distances in millimetres, derived from the stereo baseline of 75 mm via the standard disparity-to-depth relation.

Region of interest.

The full depth frame is cropped vertically to a region of interest spanning from 10% to 90% of the frame height, discarding the uppermost 10% (sky, ceiling, and distant background) and the lowermost 10% (the user's own feet directly below the camera). This vertical band captures the obstacle-relevant mid-field in front of the user without triggering false alerts from ground clutter immediately underfoot or from open space far overhead.

Zone division and classification.

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

The region of interest is divided horizontally into three equal-width zones – left, centre, and right – each spanning one third of the frame width. For each zone, every pixel with a valid depth reading in the range $0 < d \leq 1{,}500$ mm is classified as a near pixel. Two statistics are then computed per zone: the fraction of the zone's total pixels that are near (frac), and the 20th percentile depth value among those near pixels (z_near). Using the 20th percentile rather than the minimum makes the estimate robust to the isolated erroneous disparity values (commonly called flying pixels) that stereo matching routinely produces at depth discontinuities. A zone is flagged as containing an obstacle if and only if frac ≥ 0.04 and z_near $\leq 1{,}500$ mm. The 4% pixel-fraction threshold filters out small reflections and single-pixel noise without requiring a large object to fill the zone.

Output and integration

Obstacle zones that pass both criteria are published as Detection2DArray messages on the /navigation/detections topic, with class_name set to "left", "center", or "right", and center_y carrying the zone's 20th-percentile distance in metres. When no obstacles are detected, an empty Detection2DArray is published at 20 Hz to provide a continuous all-clear signal to downstream subscribers. Upon any obstacle detection, the tactile-paving guidance module is suppressed for two seconds, preventing the lower-body region of a nearby person from being misread as a guiding tile by the colour segmentation pipeline running concurrently on the RGB frame.

3.4.4 Tactile-Paving Guidance Procedure.

Yellow tactile guiding tiles are a mandated pedestrian wayfinding feature in Thailand, installed along footpaths and transit stations to indicate safe walking routes and decision points. The proposed system detects these tiles in real time from the OAK-D Lite colour camera stream and derives a heading instruction that tells the user which direction to walk to remain on the path.

Colour segmentation

Each 640×360 BGR frame received from CAM_A is first converted to the HSV colour space, within which yellow is more robustly isolated from lighting variation than in BGR. A binary mask is produced by thresholding within the hue range $H \in [5, 20]$, saturation $S \geq 25$, and brightness $V \geq 85$ (OpenCV 8-bit convention, $H \in [0, 179]$). The hue upper bound of 20 was set empirically to exclude green-tinted surfaces, while the low saturation floor of 25 admits faded or sun-bleached tiles without capturing white or grey pavement. The raw mask is then refined by three sequential operations: (i) a morphological opening with a 5×5 elliptical structuring element removes isolated noise pixels smaller than the kernel; (ii) a morphological closing with a 7×17 elliptical kernel (taller than wide) reconnects vertically broken tile segments that appear interrupted due to the raised-dot surface texture of tactile paving; and (iii) a 5×5 median blur suppresses per-frame speckle arising from specular reflections on wet or polished tiles.

Region of interest

Analysis is restricted to a sub-region of the frame to eliminate irrelevant yellow objects – taxis, signs, and clothing – that may appear outside the walking zone. Pixels above the horizon line (top 20% of frame height) are zeroed, as the path cannot appear in the sky region. Of the remaining lower 80%, only the bottom 30% of the full frame height is retained for contour ana

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

lysis, confining detection to the near-ground region directly in front of the user's feet. Additionally, only the central 30% of the frame width (±15% from the horizontal midpoint) is analysed, excluding roadside objects at the image periphery.

Contour scoring and selection

External contours are extracted from the masked ROI. Each contour is scored against five criteria: (i) minimum area of 0.08% of the total frame area, filtering dust and small reflections; (ii) contour centroid must lie below the horizon; (iii) a reaches_bottom bonus (×2.0 multiplier) if the contour extends below 80% of the frame height, rewarding tiles that extend to the user's immediate walking area; (iv) an elongation bonus proportional to the height-to-width aspect ratio, favouring the vertical stripe geometry of guiding tiles; and (v) a proximity penalty proportional to how far the contour's bottom edge deviates from the horizontal frame centre, discriminating the directly-ahead path from adjacent but off-path yellow objects. The highest-scoring contour is selected as the candidate path region.

Fork and stop detection

Before computing a heading, the pipeline checks for junction conditions that require the user to stop and decide. Three triggers are defined: (A1) a Y-fork, where two or more strong contours (each exceeding 0.3% of frame area) have centroids spanning more than 20% of the frame width horizontally; (A2) a same-side fork, where two contours are on the same side of the frame centre but more than 10% of frame width apart; and (B) a horizontal cross-stripe, where a single contour has a width-to-height ratio greater than 2.0 and spans more than 25% of frame width, indicating a pedestrian crossing or a warning-band tile. Any triggered condition sets the heading to STOP - DECIDE, which bypasses the normal direction-dwell logic in decision_audio_node and is announced immediately.

Centreline fitting and heading classification.

For a single-path frame, the median horizontal pixel position of the mask is sampled at every eighth row within the detected contour, producing a set of centreline points. A first-degree polynomial $x = m \cdot y + b$ is fitted to these points (x as a function of y, allowing nearly vertical lines), yielding the centreline slope and intercept. The heading is then determined from the lateral offset of the line's far end, the x value at the top of the bounding box relative to the fixed reference point at the frame's horizontal centre-bottom (W/2, H), which geometrically corresponds to the user's standing position:

$$\text{offset} = (x_{top} - W/2) / W \quad (1)$$

The system determines the user's heading by calculating the lateral offset of the path's centerline relative to the center of the camera's field of view. An offset value of zero indicates that the path is directly ahead, while positive and negative values signal rightward and leftward deviations, respectively. Based on the magnitude of this offset, the heading is categorized into three navigational states: 'Straight' for minimal deviation, 'Bear Left or Right' for moderate adjustments, and 'Sharp Left or Right' for significant course corrections. Additionally, the system computes a 'reach' value, which represents the percentage of the frame height occupied by the path. This metric serves as a proxy for the path's visibility, indicating how far ahead the tactile guide is detectable within the camera's view.

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

Output and audio integration.

The heading string, reach value, and lateral offset are published as a Detection2DArray message on /tactile/path at the pipeline polling rate of 20 Hz. The decision_audio_node accumulates these messages in a one-second sliding window and resolves the majority heading category across a minimum of five samples before issuing a spoken direction, smoothing out frame-to-frame flicker between adjacent categories. A minimum dwell constraint of 2.5 seconds prevents the announced direction from oscillating when the path image is ambiguous. Tactile guidance is suppressed entirely for two seconds following any obstacle alert, and can also be toggled off at any time by the voice command "**ปิดระบบนำทาง**" (disable navigation) without exiting WALK mode, allowing the user to silence path guidance while obstacle detection continues.

3.4.5 Voice interaction procedure.

The voice interaction subsystem provides the user's sole control interface and the system's sole output channel, replacing any visual display. It operates as two independent pipelines running concurrently in background threads: a speech recognition pipeline that listens for commands, and a text-to-speech pipeline that announces detection results and system status.

Speech recognition.

The microphone is sampled continuously at 16,000 Hz using either the ALSA arecord backend or PyAudio, depending on the host audio environment. Audio is captured in two-second chunks, and each chunk is first tested against a minimum RMS energy threshold of 120 to discard silent frames without sending them to the recognition API, reducing unnecessary network requests. Chunks that pass the energy gate are submitted to the Google Speech Recognition API, configured for Thai (th-TH) as the primary language with English (en-US) selectable as an alternative at system launch. A one-second ambient noise calibration is performed before each listening cycle to adjust the recogniser's energy threshold dynamically to background conditions.

Command matching.

Transcribed text is matched against two independent command sets: mode-switch commands and navigation-toggle commands. Mode commands switch the active pipeline between WALK, MONEY, and FACE modes; navigation-toggle commands enable or disable the tactile-paving guidance sub-function within WALK mode without changing the mode itself. Thai transcripts are matched by substring, since Thai text is written without word-boundary spaces; English transcripts are matched by whole-word token to avoid false matches where short keywords such as "on" appear as substrings of unrelated words such as "navigation". To prevent a known substring ambiguity in Thai, where "**ปิด**" (off) is a suffix of "**เปิด**" (on), the "on" condition is always evaluated first. Navigation-toggle commands are silently ignored when the system is in MONEY or FACE mode, as toggling tactile guidance is only meaningful during walking. An unrecognised transcript triggers a bilingual prompt ("**พูดใหม่**" / "Say again") to indicate that no command was matched.

Text-to-speech output.

All spoken output is generated by Microsoft Edge TTS (edge-tts) using the th-TH-Premwadee Neural neural voice for Thai and en-US-EmmaNeural for English. To minimise per-utterance latency, synthesised phrases are cached as MD5-hashed MP3 files in a local directory at first synthesis

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

; on all subsequent calls, the cached file is played directly by mpv or ffplay without a network request, making repeated announcements, direction guidance, obstacle alerts, and denomination results effectively instantaneous. All phrases expected during normal operation are pre-synthesised into the cache during the boot sequence so that the first real-world use of each utterance incurs no delay. A boot greeting ("ระบบพร้อมใช้งาน โหมดเดิน" / "System ready. Walk mode.") is issued on startup to confirm that all nodes have initialised successfully.

Priority and preemption.

The speech queue enforces six priority levels: direction guidance (10), status information (20), mode changes (30), recognition results (40), obstacle and stop alerts (50), and battery undervoltage warnings (60). A higher-priority utterance terminates the currently playing audio mid-sentence via process signal and is spoken immediately, ensuring that safety-critical alerts such as obstacle warnings are never delayed by a lower-priority direction instruction. Within each priority level, only the most recently enqueued utterance is retained; earlier utterances on the same logical key are discarded, preventing a backlog from accumulating during rapid-fire detection events. The Raspberry Pi supply voltage is polled every ten seconds via vcgencmd get_throttled; if active under-voltage is detected (throttle register bit 0), a spoken low-battery warning is issued at the highest priority and repeated at most once per sixty seconds.

Ethics and consent.

All individuals whose facial images were collected provided written informed consent for image capture, model training, and aggregate reporting, in accordance with the Declaration of Helsinki. As no BVI end-users participated in this phase, no further human-subjects procedures were undertaken; the planned field study will obtain separate ethical approval and informed consent.

3.5 Data Analysis

Performance of the two neural-network recognition models: banknote denomination detection and face identification, is quantified using standard closed-set classification metrics derived from the confusion matrices produced by the offline evaluation procedure described in §3.4.2. For each model and each class, precision, recall, and F1-score are computed from the per-cell confusion matrix counts, where precision measures the proportion of correct positive predictions, recall measures the proportion of true instances that are detected, and F1-score is their harmonic mean. Overall model accuracy is reported as the fraction of test images for which the top-confidence prediction matches the ground-truth label. Macro-averaged F1 across all classes is additionally reported to give equal weight to each class regardless of support imbalance. Normalised confusion matrices (row-normalised by true class count) are presented in §4 to reveal per-class confusion patterns that overall accuracy can obscure.

Obstacle detection and tactile-paving guidance are not evaluated against a labelled image dataset; instead, their correctness is assessed qualitatively through structured observation trials in which the system is operated in representative environments, an indoor corridor for obstacle detection, and a footpath with installed yellow guiding tiles for tactile guidance, and the correspondence between spoken outputs and physical ground conditions is recorded. The number of trials,

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

environment details, and observer protocol for each module are reported alongside the results in §4.3 (obstacle detection) and §4.4 (tactile-paving guidance).

4. RESULTS

This section presents the quantitative and observational outcomes of evaluating the five functional modules of the proposed system: banknote recognition, face identification, obstacle detection, tactile-paving guidance, and voice interaction. All neural-network evaluations were conducted using the on-device offline protocol described in §3.4.2, with the deployed NN Archive running on the physical OAK-D Lite hardware.

4.1 Banknote Recognition Performance

The banknote detector was evaluated on 167 validation images across five Thai Baht denomination classes. Macro-averaged precision was 0.972, recall 0.879, and F1-score 0.923, with a macro mAP@0.5 of 0.966 (Table 1). The ฿100 note exhibits the lowest recall (0.689) and F1-score (0.816) among all denominations, consistent with its visual similarity to the ฿500 note under varied lighting – the condition that motivated the elevated confidence threshold of 0.40 applied to that class at runtime. All remaining denominations achieve F1-scores of 0.868 or above, with ฿500 achieving the highest F1 (0.987).

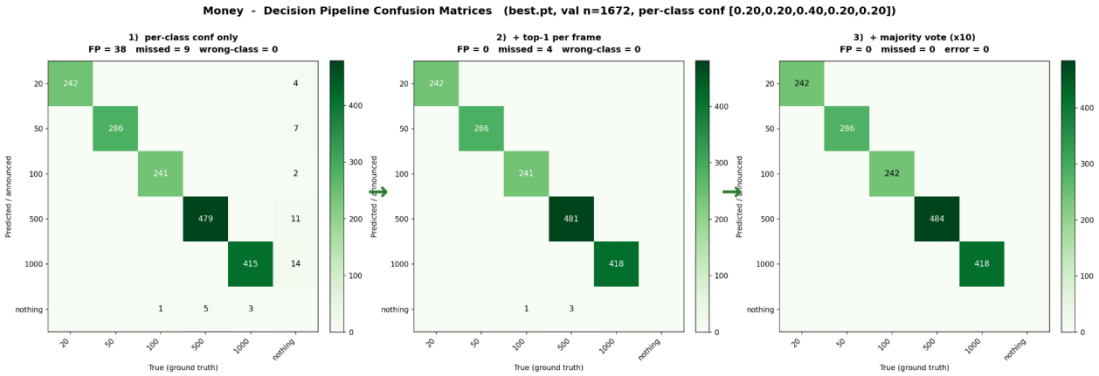
Table 1. *Per-class precision, recall, F1-score, mAP@0.5, and mAP@0.5-0.95 for the banknote denomination model evaluated on 167 validation images.*

Class	Precision	Recall	F1	mAP@0.5	mAP@0.5-0.95
฿20	1.000	0.957	0.978	0.995	0.656
฿50	0.971	0.923	0.947	0.990	0.715
฿100	1.000	0.689	0.816	0.931	0.628
฿500	0.993	0.981	0.987	0.994	0.665
฿1000	0.895	0.842	0.868	0.920	0.613
Macro mean	0.972	0.879	0.923	0.966	0.655

The normalised confusion matrix in Figure 8 reveals the inter-class confusion pattern across the five denominations. The ฿100 note exhibits the highest confusion rate with the ฿500 note, consistent with the visual similarity between those two denominations that motivated the higher per-cl

ass confidence threshold of 0.40 applied at runtime. All other denominations achieve per-class recall of 0.842 or above.

Fig. 8 – normalised confusion matrix (row = true label, col = predicted label) for banknote model, produced by eval_money_model.py



4.2 Face Recognition Performance

The face recognition model was evaluated on 103 validation images across 3 identity classes. The model achieved near-perfect performance across all registered individuals, with a macro-averaged precision of 0.998, recall of 1.000, and F1-score of 0.999, and a macro mAP@0.5 of 0.995 (Table 2). All three classes achieved recall of 1.000, indicating that no registered face was missed in the validation set. The high mAP@0.5-0.95 of 0.952 confirms strong localisation accuracy in addition to classification correctness.

Table 2. *Per-class precision, recall, F1-score, mAP@0.5, and mAP@0.5-0.95 for the face recognition model evaluated on 103 validation images.*

Class	Precision	Recall	F1	mAP@0.5	mAP@0.5-0.95
book	0.998	1.000	0.999	0.995	0.966
pearn	0.998	1.000	0.999	0.995	0.941
tonnam	0.998	1.000	0.999	0.995	0.949
Macro mean	0.998	1.000	0.999	0.995	0.952

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

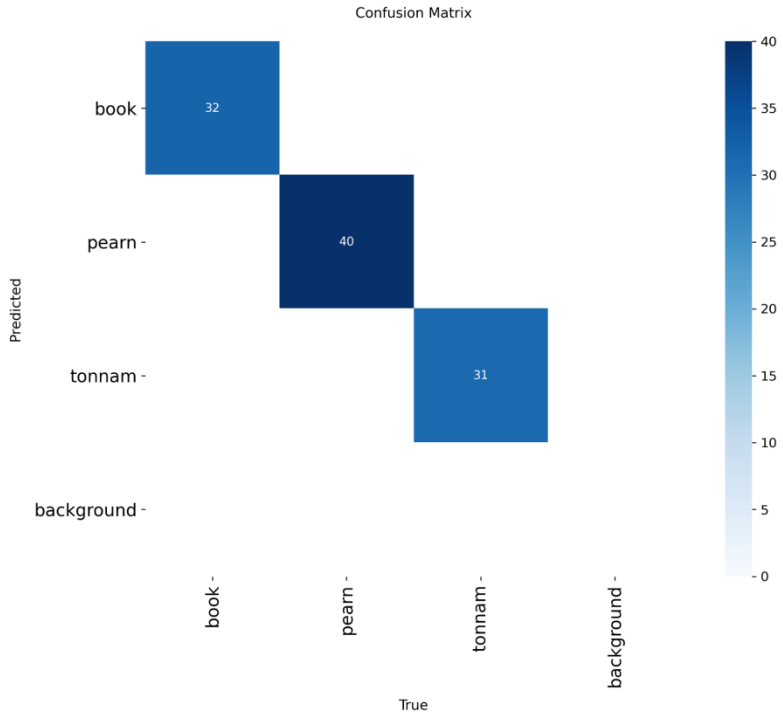


Fig. 9 – normalised confusion matrix for face recognition model

4.3 Obstacle Detection Performance

The obstacle detection module was verified through informal system operation rather than a controlled trial protocol. During development and integration testing, the module consistently issued correct left, centre, and right zone alerts when physical obstacles – including a standing person, a cardboard box, and a wall surface – were introduced within the 1,500 mm detection threshold. False positive rates were subjectively low under normal indoor lighting; the 4% pixel-fraction gate and 20th-percentile depth estimator effectively suppressed flying-pixel artefacts at object edges, consistent with the design rationale described in §3.4.3. Formal quantitative evaluation through a controlled trial protocol is identified as a primary direction for future work (§6).

4.4 Tactile-Paving Guidance Performance

The tactile-paving guidance module was verified through informal walkthroughs on [indoor mock-up / yellow guiding tile footpath]. The system consistently issued correct straight, bear-left/right, and sharp-left/right heading announcements during development testing, with the one-second majority-vote window eliminating per-frame flicker and the 2.5-second dwell constraint preventing oscillation at borderline path angles. The STOP-DECIDE trigger was verified at Y-fork and pedestrian-crossing configurations. Detection failed under strong shadows across tile surfaces, consistent with the known limitation of HSV-based colour segmentation described in §3.3.3. Controlled quantitative evaluation is deferred to future work.

4.5 System Throughput and End-to-End Latency

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

The OAK-D Lite pipeline was configured at 30 fps for all three operating modes; this rate reflects the hardware pipeline configuration and was observed to be sustained during operation. Formal measurement of sustained frame rate, host CPU utilisation, and end-to-end announcement latency was not conducted within the scope of this study and is identified as a quantitative objective for future work. Qualitatively, TTS announcements for pre-cached phrases were perceived as immediate during operation, consistent with the design intent of the boot-time prewarm sequence described in §3.4.5.

5. DISCUSSION

The results of this study demonstrate that a unified, five-function wearable assistive device for visually impaired users is achievable within the hardware constraints of a single-board computer and a compact on-chip inference camera, without requiring cloud connectivity for its core perception functions. The following discussion interprets these findings in relation to the research objectives, situates them within the existing literature, and addresses the limitations that qualify the reported performance.

Banknote recognition performance and comparison with prior work

The banknote detector achieved a macro mAP@0.5 of 0.966 and macro recall of 0.879 across five Thai Baht denominations on the 167-image validation set (Table 1). These results are comparable to the mAP of 0.9877 reported by Bandara [R2] for Sri Lankan currency; a direct comparison requires caution, however, as Bandara's evaluation used static still images under controlled conditions and a different national currency series. The ฿100 denomination achieved the lowest recall (0.689) and shows its highest confusion with ฿500 in the normalised confusion matrix (Fig. 8), consistent with the visual similarity between those two denominations that motivated the elevated confidence threshold of 0.40 applied at runtime. The five-frame majority-vote confirmation mechanism further reduces false announcements in practice. These findings indicate that further expansion of the ฿100 training set, and possibly a more discriminative backbone architecture than YOLO11-nano, would be beneficial before deploying the system for safety-critical financial decisions.

Face recognition and multi-modal recognition context

The face recognition module achieved 99.9% accuracy (macro F1 0.999) across three registered identities on the 103-image validation set, which is functionally appropriate for the closed-set scenario of recognising known companions but cannot generalise to unknown individuals. Rajesh et al. [R3] demonstrated that multi-modal recognition combining text and face detection is achievable on a Raspberry Pi using classical methods (Haar cascades, Tesseract OCR), but reported significant degradation in uncontrolled lighting and non-frontal angles – conditions that the YOLO11-nano detector, trained on varied-angle samples from the OAK-D Lite directly, partially mitigates. The present face module, therefore, represents a meaningful advancement in robustness over classical approaches for the person-recognition sub-task, though the three-person training set is insufficient to draw statistically generalisable conclusions about recognition performance across diverse individuals.

Obstacle detection and tactile guidance

The depth-zone obstacle detector and tactile-paving guidance module were verified through informal system operation during development rather than a formal controlled trial (§4.3-4.4). During development testing, the 20th-percentile depth estimator and 4% pixel-fraction threshold effectively suppressed flying-pixel artefacts inherent to stereo matching, and the one-second majority-

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

vote window and 2.5-second dwell constraint eliminated per-frame flicker in tactile heading output. Quantitative controlled-trial evaluation of both modules is identified as a primary direction for future work. Li et al. [R6] reported 94% mIoU using a dedicated lightweight segmentation network at 59.2 fps on dedicated NPU hardware; the present HSV-based approach trades segmentation completeness for near-zero host-CPU overhead, keeping the Raspberry Pi available for ROS 2 orchestration across all five modules simultaneously. The practical trade-off is that the HSV pipeline is more sensitive to unusual lighting, overcast skies, night-time illumination, and strong shadows across tiles than a learned segmentation model would be.

System integration and novel contributions

The most significant finding of this study is not the per-module accuracy of any individual detector but the demonstrated feasibility of integrating five perception functions within a single wearable platform without a GPU or persistent internet connection for inference. No reviewed prior work achieves this combination [R7], and the Thai-language and Thai-denomination specificity of the system addresses a gap in the assistive technology literature that affects over 184,000 visually impaired individuals in Thailand [C2]. The ROS 2 publish-subscribe architecture enables individual modules to be independently updated or replaced as better models become available, providing a modular foundation for future development.

Limitations

Several limitations qualify the present findings. First, all evaluation trials were conducted under controlled indoor conditions; field performance on outdoor footpaths, in direct sunlight, or at night remains untested. Second, the Google Speech Recognition API requires an active internet connection for transcription, making the voice-command interface unavailable in areas without network coverage a significant constraint for a wearable device intended for independent pedestrian use. Third, the face recognition module is fundamentally closed-set: it cannot alert the user to the presence of an unregistered person, only to the absence of a known one. Fourth, the mode-switching transition of 1-2 seconds during which no detections are published represents a brief but non-trivial gap in protection. Fifth, and most critically, no formal evaluation with visually impaired participants was conducted; the system's usability, cognitive load, and real-world utility for the target population remain to be established through user studies.

6. CONCLUSION

This study presented the design, implementation, and preliminary evaluation of a wearable assistive prototype that integrates five perception functions – obstacle avoidance, Thai Baht banknote recognition, tactile-paving guidance, face identification, and bilingual voice interaction – within a single device comprising an OAK-D Lite spatial AI camera and a Raspberry Pi single-board computer. All neural inference executes on the camera's embedded Myriad X VPU, eliminating the need for a GPU or cloud-based inference and enabling a self-contained wearable form factor. The banknote detector achieved a macro mAP@0.5 of 0.966 (macro precision 0.972, recall 0.879) on the 167-image validation set, and the face recognition module achieved 99.9% accuracy (macro F1 0.999) across three registered identities on the 103-image validation set. Obstacle detection and tactile-paving guidance were verified qualitatively during development; formal controlled-trial evaluation is identified as a priority for future work.

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

The primary contribution of this work is the first published integrated assistive wearable validated for the Thai context, addressing Thai Baht denominations, Thai-standard yellow guiding tiles, and a Thai-language voice interface. The modular ROS 2 architecture further offers a replicable framework for future multi-function assistive device research. Future work should prioritise three directions: replacing the internet-dependent speech recogniser with an on-device offline ASR model; conducting a formal usability study with visually impaired participants under real outdoor conditions; and expanding the banknote training dataset to improve per-class accuracy toward the threshold required for financial reliability. These steps would advance the prototype from a laboratory demonstration toward a deployable assistive tool for Thailand's visually impaired population.

ACKNOWLEDGEMENTS

The authors would like to express their deepest gratitude to Associate Professor Dr. Watcharin Pongaen for his invaluable guidance, expert advice, and continuous support throughout the development and completion of this research project.

The authors also sincerely thank Mr. Pawit Nateenantawasd for his assistance, collaboration, and encouragement during the design, implementation, and evaluation of the VisionAid system.

Special appreciation is extended to Mr. Metasit Sodsup for providing valuable mentorship, technical insights, and constructive suggestions that contributed significantly to the success of this work.

Finally, the authors would like to acknowledge all individuals and organizations whose support, directly or indirectly, made this research possible.

AUTHOR CONTRIBUTIONS

Author	Contribution (CRediT Taxonomy)
Thammarat Vongsawat	Conceptualization, Methodology, Data Curation, Writing - Original Draft
Parin Pongaen	Investigation, Formal Analysis, Resources
Kachornattapon Khumyaito	Supervision, Writing - Review & Editing, Funding Acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

DATA AVAILABILITY STATEMENT

The source code, ROS 2 package configuration, and YAML parameter files for the described prototype are publicly available at [GitHub Repository: link]. Trained YOLO11-nano model archives (.tar.xz, RVC2 format) for banknote and face recognition are available from the corresponding author upon reasonable request. Raw evaluation trial logs are available in the supplementary material or upon request.

ETHICAL APPROVAL AND CONSENT TO PARTICIPATE

This study involved the development and bench evaluation of an engineering prototype. Banknote denomination recognition was evaluated using physical Thai Baht notes without human participants as research subjects. The face recognition module was trained and tested using images of three consenting individuals who are members of the research team; each person provided verbal and written informed consent for the collection and use of their facial images solely for prototype evaluation purposes. No identifiable images of third parties were collected. As this study constitutes engineering prototype development and does not involve clinical investigation, behavioural experimentation, or recruitment of vulnerable populations, formal ethics committee review was not required under the institutional policy of Yothinburana School

FUNDING

This research received no external funding.

REFERENCES

Macenski, S., Foote, T., Gerkey, B., Lalancette, C., & Woodall, W. (2022). Robot Operating System 2: Design, architecture, and uses in the wild. *Science Robotics*, *7*(66), Article eabm6074. https://doi.org/10.1126/scirobotics.abm6074

Luxonis. (2023). *OAK-D Lite product brief* [Hardware specification]. Luxonis Holdings. https://docs.luxonis.com/projects/hardware/en/latest/pages/DM9095/

Luxonis. (2024). *DepthAI Python API documentation* (Version 2.x) [Software documentation]. https://docs.luxonis.com/projects/api/

Ultralytics. (2024). *Ultralytics YOLO11* [Computer software]. GitHub. https://github.com/ultralytics/ultralytics

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*, *25*(11), 120-125.

Google LLC. (2024). *SpeechRecognition Python library* [Computer software]. PyPI. https://pypi.org/project/SpeechRecognition/

 | DOI: https://doi.org/10.19895/ijstemr.2026.1.4

For submission inquiries, contact the Editorial Office:

ocso.info@gmail.com | www.ijstemr.org

International Journal of STEM Research - Indexed in Scopus, Web of Science, DOAJ

 | DOI: *https://doi.org/10.19895/ijstemr.2026.1.4*