International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(9s) 2026 pp. 50-71
https://doi.org/10.70917/ijcisim-2026-3422
Article

Empirical Evaluation of a Fully Offline, Low-Cost Edge AI Wearable Smart Navigation Assistant for Distance-Aware Obstacle Detection

Shubhangi Rajkumar Katke[1,2], Utkarsha Pacharaney

[1]Department of Artificial Intelligence and Data Science, Datta Meghe College of Engineering, Mumbai, India, ORCID: 0009-0007-0730-601X. shubhangi.katke@dmce.ac.in
[2] Faculty of Engineering and Technology, Datta Meghe Institute of Higher Education and Research (DU), Wardha, Maharashtra, India, utkarshap.feat@dmiher.edu.in

Abstract: Real-time visual perception on resource-constrained embedded hardware must reconcile computational economy, low latency, and dependable sensing accuracy within tight power and cost envelopes. This paper reports on the Smart Navigation Device (SND), a wearable assistive perception system that performs object detection, distance ranging, sensor fusion, and speech feedback entirely on a Raspberry Pi 4B without any cloud dependency. At the core of the system is the Cascaded Detection-Ranging Fusion (CDRF) framework, a four-stage pipeline that couples the lightweight YOLO11n detector with ultrasonic time-of-flight ranging through confidence-guided detection acceptance, spatial-zone partitioning, dominant-object association, and adaptive suppression of redundant announcements. The hardware-software co-design keeps every processing stage — image capture, neural inference, ranging, fusion, and text-to-speech synthesis — local to the device, eliminating transmission latency, removing a major privacy exposure, and preserving operability where network connectivity is unreliable or absent. The framework was evaluated across 169 controlled trials spanning three obstacle categories — person, chair, and laptop — at distances from 0.30 m to 4.20 m. Detection rates of 90.4%, 92.9%, and 77.0% were obtained for the three classes respectively, with mean absolute ranging errors of 2.11 cm, 1.60 cm, and 1.79 cm. Agreement between the ultrasonic estimate and ground-truth distance was excellent (Pearson $r = 0.9995$, $p < 10^{-219}$), and Bland–Altman analysis revealed a small systematic bias of −1.10 cm (95% limits of agreement: −7.88 cm to 5.68 cm). A chi-square test indicated a statistically meaningful distance-dependent decline in laptop-class detection reliability. Benchmarked against previously reported wearable travel aids, the SND achieves comparable or better detection reliability than low-cost ultrasonic-only alternatives while additionally providing object identity, and does so at a fraction of the hardware cost and without any of the connectivity dependencies of cloud-assisted alternatives — positioning it as a reproducible, statistically grounded, and economically accessible baseline for future assistive-perception research.

Keywords: Real-time image processing, embedded vision, edge AI, YOLO11n, sensor fusion, Raspberry Pi, object detection, distance estimation, wearable perception, assistive navigation, accessibility engineering, human-centred design

1. INTRODUCTION

Visual impairment affects an estimated 2.2 billion people worldwide, of whom approximately 43 million live with blindness [1]. The loss of navigational independence that accompanies severe vision loss carries direct consequences for employment, education, and social participation [2], and longitudinal epidemiological projections indicate that the global burden of vision impairment continues to grow as populations age [3]. The white cane remains the most widely deployed mobility aid because of its simplicity and reliability, but it conveys proximity through tactile contact alone and provides no information about what an obstacle actually is, which limits a user's ability to distinguish a genuine hazard from a harmless object in the travel path [4].

Electronic travel aids (ETAs) that incorporate computer vision have long been proposed as a richer alternative to the cane, promising both obstacle detection and obstacle recognition in a single device [5]. In practice, however, the dominant commercial paradigm has tended toward one of two extremes: inexpensive ultrasonic-only aids that detect proximity but cannot identify objects, or expensive, cloud-assisted smart-glasses platforms that identify objects richly but depend on a live network connection and a recurring subscription or API cost [6], [7]. Both extremes create adoption barriers. The first sacrifices situational understanding; the second sacrifices affordability and reliability in precisely the low-connectivity or infrastructure-limited environments where an assistive device is most needed. This tension is not a new observation — ultrasonic obstacle-avoidance aids for mobility have been studied since at least the 1980s [32] — but it has become more consequential as vision-based ETAs have shifted the state of the art toward cloud-dependent architectures.

Access to assistive technology is, moreover, not merely a matter of consumer convenience. The World Health Organization's International Classification of Functioning, Disability and Health frames environmental barriers — including the unavailability of appropriate assistive products — as a direct determinant of disability outcomes rather than an incidental inconvenience [48], and the United Nations Convention on the Rights of Persons with Disabilities explicitly identifies accessibility, including access to assistive technologies, as a precondition for the exercise of independence and full participation in society [47]. Framed this way, the cost and connectivity barriers described above are not simply engineering shortcomings; they are barriers to a basic form of autonomy for a substantial fraction of the global population, and they motivate a design philosophy in which affordability and offline operability are treated as first-class requirements rather than secondary optimizations.

Recent advances in lightweight neural network architectures — in particular the YOLO11 family of object detectors [8], [9] — have made real-time, CPU-only object detection practical on single-board computers priced for consumer use. When combined with ultrasonic time-of-flight ranging, which adds negligible hardware cost, these advances create the conditions for a wearable perception system that runs entirely offline: one that does not degrade in poor-connectivity environments and does not transmit a continuous video feed of a user's surroundings to a remote server. This offline property is not only a cost consideration. In privacy-sensitive and clinical deployment contexts, the complete absence of a network dependency is increasingly treated as an explicit design and regulatory requirement rather than a nice-to-have feature [10].

Despite the intuitive appeal of such systems, the published literature on low-cost wearable ETAs contains a recurring empirical gap. Many prototype-stage contributions report system architecture and a qualitative demonstration of functioning hardware without controlled, distance-stratified, trial-level accuracy or ranging data [11], [12]. Where quantitative results are reported, they frequently take the form of a single aggregate accuracy figure, which conceals distance-dependent and object-size-dependent degradation effects that matter a great deal for real deployment — a device that works reliably at arm's length but fails at three metres behaves very differently in practice than the aggregate number suggests. Statistical instruments that are standard in other measurement-validation contexts, such as Bland–Altman agreement analysis [44] and formal significance testing of accuracy across strata [46], are correspondingly rare in this specific sub-field, even though they are inexpensive to apply once trial-level data exists.

This paper addresses both gaps directly. Its contributions are fourfold:

1) the Cascaded Detection–Ranging Fusion (CDRF) framework, a formally specified, single-sensor fusion and decision architecture that couples a nano-scale object detector with a single ultrasonic ranging sensor;

2) A complete mathematical formulation of detection confidence, spatial-zone assignment, ranging-uncertainty propagation, and announcement-suppression logic;

3) A controlled, 169-trial empirical evaluation across three object classes and four distance bands, reporting detection rates with confidence intervals, ranging accuracy (MAE, RMSE, percentage error), Bland–Altman agreement, and chi-square significance testing; and

4) A literature-grounded comparative evaluation that positions the measured performance of the Smart Navigation Device (SND) against previously reported wearable and embedded travel aids on the axes of sensing modality, offline capability, hardware cost, and validation rigor.

The remainder of this paper is organized as follows. Section 2 reviews related work in edge-deployed object detection, assistive navigation systems, and sensor fusion for proximity estimation. Section 3 presents the system architecture and the CDRF framework in full mathematical detail, with symbol definitions collected in the Nomenclature table preceding this Introduction. Section 4 describes the experimental protocol. Section 5 reports detection and ranging results with statistical analysis. Section 6 benchmarks these results against prior work. Section 7 discusses the findings in depth, including human-factors and deployment considerations. Section 8 sets out a structured, near- to long-term future scope, and Section 9 concludes the paper.

2. RELATED WORK

2.1 Object Detection Architectures for Edge Deployment

The evolution from two-stage detectors — Faster R-CNN [13] and its successors — toward single-shot architectures reflects a long-running trade-off between accuracy and inference latency. SSD [14] and the original YOLO formulation [15], [16] resolved this tension by treating detection as a direct regression over a spatial feature grid rather than a two-stage propose-then-classify pipeline, which made real-time inference achievable on mid-range hardware for the first time. Successive YOLO generations pushed this frontier further: YOLOv4 [17] consolidated a set of training and architectural refinements into a single competitive model, and YOLOv7 [18] introduced trainable bag-of-freebies strategies that improved the accuracy–latency frontier specifically for embedded and edge targets, while YOLOv6 [19] targeted industrial deployment with a re-parameterized backbone optimized for inference throughput. YOLO11n [9], the nano variant of the YOLO11 generation used in this work, compresses the backbone and neck further through depthwise convolutions and lightweight attention modules, achieving competitive mean Average Precision on MS-COCO [20] at a parameter count compatible with CPU-only inference. Independent benchmark studies of YOLO-family models on Raspberry Pi-class hardware confirm that YOLO11n offers a favourable latency profile among current-generation variants on the ARM Cortex-A72 core [21], and broader surveys of the object-

detection field over the past two decades situate this nano-scale, single-shot design lineage as the practical endpoint for CPU-bound embedded inference [22].

An alternative lightweight lineage, built around depthwise-separable convolutional backbones such as MobileNet [23] and MobileNetV2 [24], and compound-scaled detectors such as EfficientDet [25], offers a comparable latency–accuracy trade-off and is frequently combined with post-training quantization [26] to further reduce inference cost on integer-only edge accelerators. These architectures were considered during system design; YOLO11n was ultimately selected on the basis of its native single-shot detection head, its maturity within an actively maintained open-source toolchain, and empirically observed inference latency on the target Raspberry Pi 4B hardware that matched the real-time requirement of the CDRF pipeline described in Section 3 without requiring a separate quantization or conversion step.

2.2 Assistive Navigation Systems: State of the Art

Wearable vision assistants for people who are blind or have low vision have been reviewed extensively [4], [11]. Early electronic canes augmented tactile feedback with ultrasonic proximity warning [32] but could not identify the obstacle they detected, only its presence and rough range. The deep learning era enabled CNN-based wearable systems that close this gap: Bai . demonstrated a CNN-based indoor object-recognition system for visually impaired users with audio output, reporting high recognition accuracy in controlled indoor settings [28]. At the higher end of the market, Stramba-Badiale . conducted a clinical pilot evaluation of the OrCam MyEye device and reported strong performance on text-reading and face-recognition tasks [29]; the device performs its core recognition on-device, but its retail cost remains prohibitive for a large share of the population that would benefit most, particularly in low- and middle-income settings [6]. Commercial smart-glasses platforms that integrate large vision-language models achieve impressive open-ended scene understanding [7], but this capability is presently delivered through sustained cloud connectivity with recurring inference costs, which disqualifies these platforms for offline or infrastructure-limited deployment. A systematic review by Bhatlawande . of prototype obstacle-avoidance systems identified affordability and connectivity independence as the two most frequently cited barriers to real-world adoption of assistive ETAs, particularly in resource-constrained settings, directly motivating the design priorities adopted in this work [30].

2.3 Sensor Fusion for Proximity and Depth Estimation

Ultrasonic time-of-flight sensing remains the lowest-cost mechanism available for single-axis proximity estimation, and its behaviour is well characterised both by manufacturer specification [31] and by decades of robotics literature on sonar-based ranging and mapping, beginning with early foundational work on sonar-based world modelling [33] and ultrasonic obstacle avoidance [32]. Its principal weaknesses — single-beam angular ambiguity and specular-reflection artefacts from smooth or angled surfaces — are likewise well documented in that literature.

Infrared proximity sensing offers a comparable low-cost single-axis alternative with different failure modes, particularly under variable ambient lighting [34]. Stereo vision and structured-light depth sensors extend proximity sensing to full spatial coverage at higher computational and hardware cost; RGB-D-based navigation-assistance systems for visually impaired users have demonstrated the value of dense depth information for obstacle avoidance, at the cost of sensor bulk and power draw that complicate continuous wearable use [36]. LiDAR-based wearable navigation systems, reviewed by Tapu ., offer dense, high-precision depth maps but at a hardware cost an order of magnitude above the sub-USD-50 target adopted in this work [35]. Monocular depth-estimation networks — including MiDaS [37] and the more recent Depth Anything V2 [38] — infer dense depth from a single RGB frame using large-scale self-supervised training, and represent a technically compelling future successor to single-beam ultrasonic ranging; at present, however, their CPU inference latency on Raspberry Pi 4B-class hardware exceeds the real-time threshold required for a

continuous wearable perception loop, which motivates the single-sensor ultrasonic approach adopted here while identifying monocular depth as the primary migration target for future hardware generations (Section 8). Multi-object tracking methods such as SORT [40], its deep-association extension DeepSORT [41], and the more recent ByteTrack [39] extend single-frame detections into temporal object trajectories and are directly applicable to strengthening the announcement-suppression mechanism described in Section 3.2.4; more broadly, the general challenges of combining heterogeneous sensor modalities with different noise characteristics and update rates are catalogued in the multisensor data fusion literature [43], which frames the design choices made in the CDRF fusion stage within a wider body of established fusion practice.

2.4 Positioning of the Present Work

Taken together, the reviewed literature suggests that the design space for wearable ETAs is defined by at least three requirements that are rarely satisfied simultaneously by a single system: full offline operation, combined obstacle identity and distance output, and a hardware cost low enough for wide accessibility. Ultrasonic-only canes satisfy the first and third requirements but not the second; cloud-assisted smart glasses satisfy the second but not the first or third; clinically validated on-device systems such as OrCam MyEye satisfy the first and second but not the third. The Smart Navigation Device, described in Section 3, is designed to satisfy all three simultaneously. Its distinguishing characteristic relative to prior low-cost prototype papers, however, is not the design goal itself — several previous works share similar ambitions, as reviewed above — but the presentation of a formally specified fusion framework accompanied by controlled, distance-stratified, trial-level empirical data with agreement and significance analysis, rather than a qualitative demonstration alone. Section 6 returns to this positioning with a direct, literature-grounded comparison once the empirical results of Section 5 have been established.

3. SYSTEM METHODOLOGY

This section presents the SND hardware architecture and the Cascaded Detection–Ranging Fusion (CDRF) framework that governs its perception-to-feedback pipeline. Symbols introduced below are collected, with units, in the Nomenclature table at the head of this document.

3.1 Hardware Architecture

The SND prototype is built on a Raspberry Pi 4B (Broadcom BCM2711, quad-core ARM Cortex-A72 at 1.5 GHz, 4 GB LPDDR4 RAM), chosen for its balance of CPU performance, GPIO availability, and unit price. A USB HD camera captures frames at 320 × 240 px in MJPEG format; this resolution was selected empirically as the lowest at which YOLO11n maintains consistent class separation across the tested object set, minimising capture and decode latency T_c without sacrificing detection reliability. An HC-SR04 ultrasonic sensor, interfaced through two dedicated GPIO pins (trigger and echo), provides single-axis distance estimates along the device's forward axis. All audio feedback is synthesised locally by the eSpeak-NG text-to-speech engine and dispatched through an asynchronous priority queue, which prevents audio playback from blocking the main detection loop. No component of the perception-to-feedback pipeline requires, initiates, or benefits from any network connection; the system behaves identically regardless of connectivity state, which is the central design property motivated in Section 1.

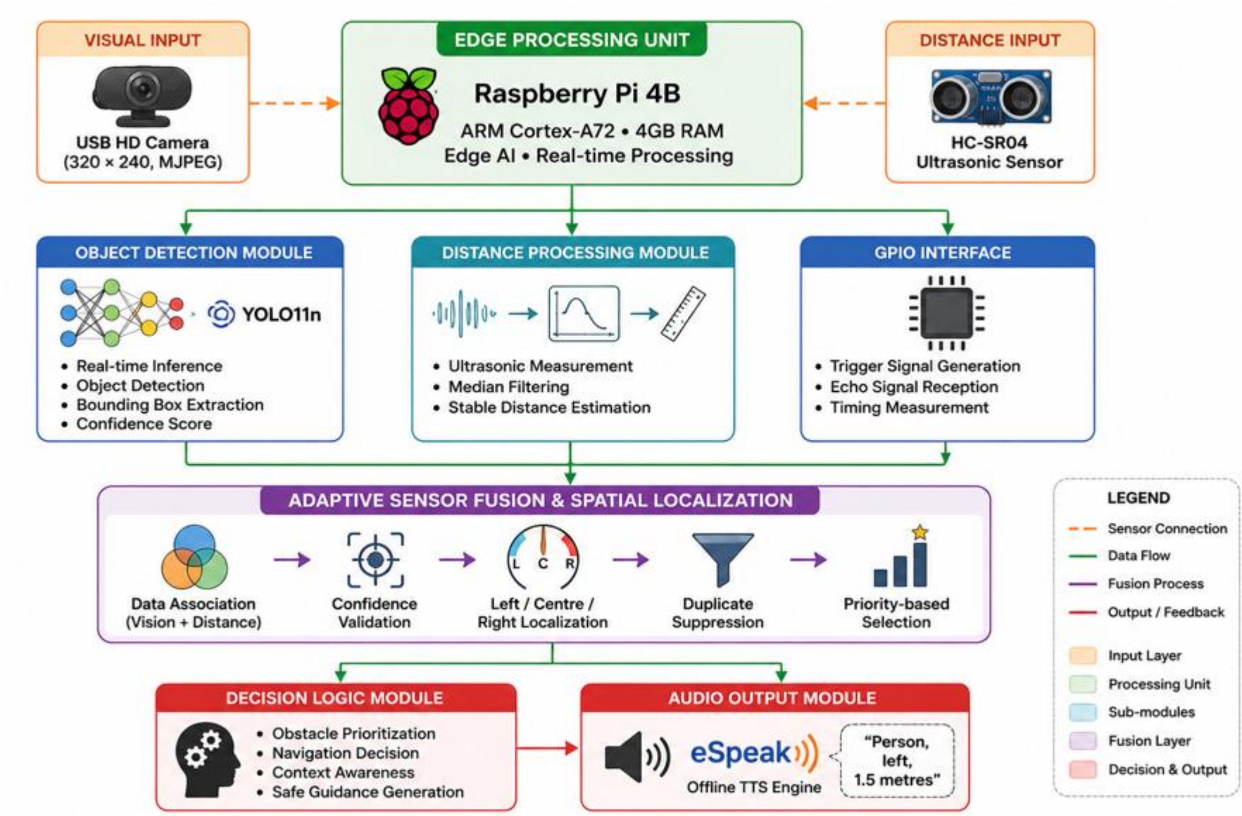

Figure 1: System Architecture of the Smart Navigation Device (SND).

Fig. 1 depicts the embedded processing architecture executed on the Raspberry Pi. Visual information acquired from the USB camera is processed using the lightweight YOLO11n object detection network, whereas obstacle distance is simultaneously measured using the ultrasonic sensor through the GPIO interface. A median filtering stage reduces measurement noise before the visual and ranging information are integrated by the sensor fusion module. The fusion process estimates object identity, relative position (Left, Centre, or Right), and obstacle distance, after which a decision module suppresses redundant announcements and forwards only relevant navigation information to the offline eSpeak text-to-speech engine.

3.1.1 Physical Prototype and Field Demonstration

Beyond the bench-style, tape-measure-controlled trials described in Section 4, the assembled SND prototype was also exercised in an informal, hand-held field demonstration to confirm that the complete perception-to-speech loop — capture, YOLO11n inference, HC-SR04 ranging, CDRF fusion, and eSpeak-NG audio announcement — operates correctly outside the fixed bench rig. Fig. 2 reproduces representative frames from this demonstration, showing the assembled hardware (camera, HC-SR04 module, and Raspberry Pi 4B wired on a breadboard prototype) and the device in operation as it is carried through an indoor scene containing everyday obstacles, including a person, consistent with the object classes evaluated quantitatively in Section 5. These frames are presented as qualitative, illustrative evidence that the architecture of Fig. 1 corresponds to a physically realised, working device rather than a simulation; they are not a substitute for the controlled, distance-stratified trial data of Sections 4–5, which remains the quantitative basis for every accuracy and ranging figure reported in this paper.

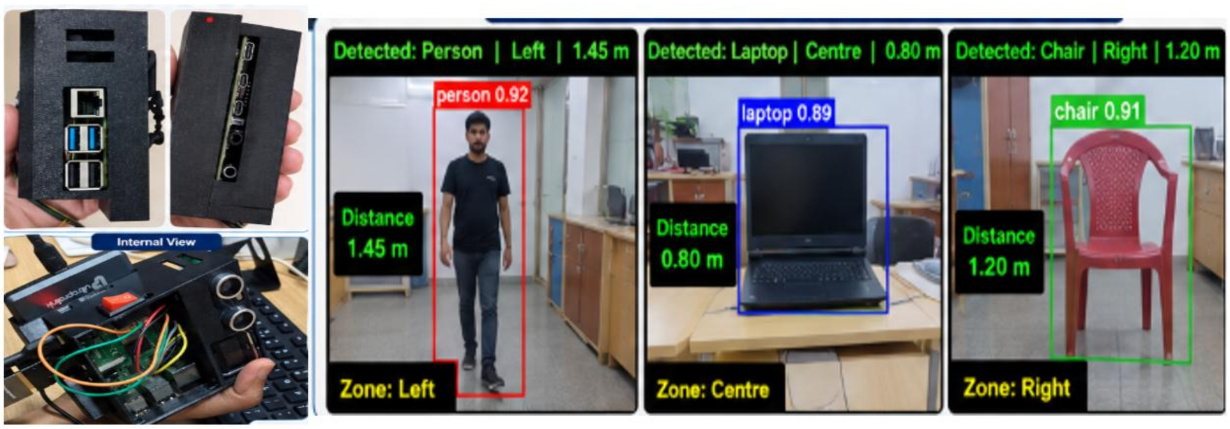

Figure 2: The assembled SND prototype: hardware close-up and live indoor operation real time output on Raspberry pi video window (frames extracted from the accompanying demonstration recordings).

The fabricated hardware prototype is shown in the left section of Fig. 2, including the custom enclosure housing a Raspberry Pi 4B, a USB HD camera, an HC-SR04 ultrasonic sensor, a rechargeable power source, and supporting electronic interfaces. The compact design enables portable operation while keeping hardware costs low and deployment easy. The other part demonstrates successful detection of common indoor objects such as a person, a chair, and a laptop. Each detected object is enclosed in a bounding box, along with its confidence score, estimated distance, and spatial direction. These experimental results confirm that the proposed system can simultaneously perform object recognition, obstacle ranging, and directional localisation using low-cost embedded hardware while operating entirely in offline mode.

3.2 The Cascaded Detection-Ranging Fusion (CDRF) Framework

The CDRF framework specifies the complete processing pipeline from frame acquisition to audio announcement as a four-stage cascade: (1) Detection, (2) Ranging, (3) Fusion, and (4) Decision. Each stage is formally specified below; the complete execution flow is illustrated in Fig.3. The software workflow describes the sequential execution of the proposed algorithm. After system initialisation, the application repeatedly captures image frames, acquires ultrasonic measurements in parallel, performs object detection, computes object location, estimates obstacle distance, executes the fusion algorithm, filters duplicate detections, and generates spoken navigation instructions. This continuous processing loop enables real-time environmental awareness without requiring cloud communication or external computational resources.

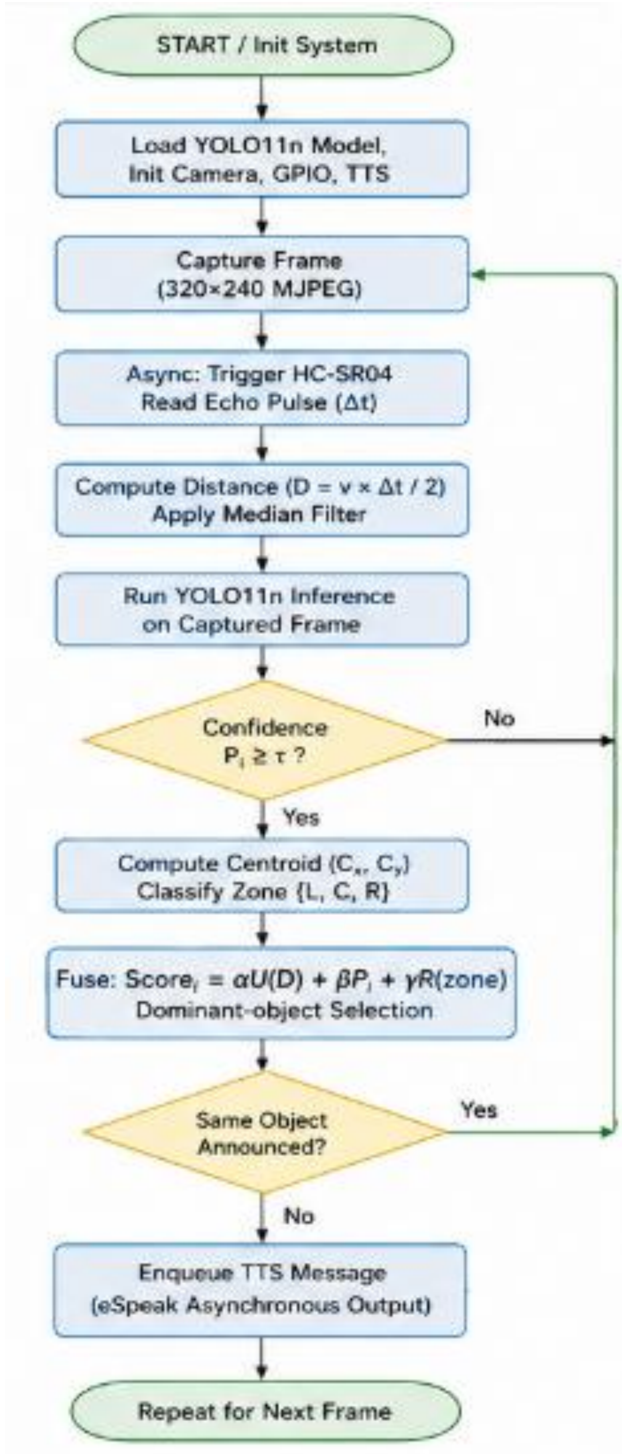

Figure 3: CDRF software execution flowchart. Diamond nodes represent decision gates; the outer loop runs continuously.

3.2.1 Stage 1: Detection

Each captured frame I(x, y, t) at time index t is passed to the YOLO11n inference engine, which outputs a set of candidate detections:

$$D(I) = \{ (B_i, c_i, p_i) \mid i = 1...N \} \quad (1)$$

where $B_i = (x_i, y_i, w_i, h_i)$ is the axis-aligned bounding box, $c_i \in C$ is the predicted class label, and p_i is the composite confidence score:

$$p_i = P(\text{object} \mid I, B_i) \times P(c_i \mid \text{object}, I, B_i) \quad (2)$$

A detection is accepted by the CDRF framework if and only if its confidence exceeds a tunable threshold τ:

$$D_\tau = \{ (B_i, c_i, p_i) \in D(I) : p_i \geq \tau \} \quad (3)$$

The bounding-box centroid and normalised area — the two geometric features used downstream — are:

$$C_{x,i} = x_i + w_i/2, \quad C_{y,i} = y_i + h_i/2 \quad (4)$$

$$\hat{A}_i = (w_i \times h_i) / (W \times H) \quad (5)$$

The normalised area $\hat{A}_i$ serves as a proxy for relative proximity: for a fixed object at distance d, its pixel footprint scales approximately as $\hat{A}_i \propto d^{-2}$, a consequence of the inverse-square relationship between subtended solid angle and distance under perspective projection. This relationship is revisited in Section 7.1 to explain the class-dependent detection degradation observed at range.

3.2.2 Stage 2: Ranging

The HC-SR04 sensor estimates distance from the round-trip travel time t_{echo} of a 40 kHz ultrasonic pulse:

$$D_{raw} = v(T_{amb}) \times t_{echo} / 2 \quad (6)$$

where the speed of sound v, corrected for ambient temperature T_{amb} (°C), is:

$$v(T_{amb}) \approx 331.3 + 0.606 \times T_{amb} \; [\text{m s}^{-1}] \quad (7)$$

Single-shot HC-SR04 readings contain impulsive noise arising from specular echoes and multi-path reflections. The CDRF framework applies a sliding-window median filter of length k over consecutive readings $D_1...D_k$ to obtain a robust estimate:

$$D_m = \text{median}(D_1, D_2, ..., D_k) \quad (8)$$

The ranging uncertainty — defined as the standard deviation of the filtered estimate across repeated measurements of a static target at distance d — grows with d due to increasing beam-divergence-related cross-sectional overlap with background surfaces:

$$\sigma_D(d) \approx \alpha_0 + \alpha_1 \times d \quad (9)$$

where α_0 and α_1 are empirically calibrated, sensor-dependent constants. This model motivates the distance-dependent growth in MAE documented in Section 5.3.

3.2.3 Stage 3: Fusion

Because the SND employs a single forward-facing ultrasonic sensor, its scalar range reading must be attributed to the detection most likely to coincide with the sensor's beam axis. The CDRF fusion rule selects the dominant detection i* as the accepted detection with the maximum normalised area within the centre spatial zone:

$$i^* = \text{argmax}_i \left[\hat{A}_i \times \mathbb{1}(L_i = \text{Centre}) \right] \quad (10)$$

where the spatial zone label L_i is assigned by partitioning the frame width W into three equal horizontal zones:

$$L_i = \{ \text{Left if } C_{x,i} < W/3; \text{ Centre if } W/3 \leq C_{x,i} \leq 2W/3; \text{ Right if } C_{x,i} > 2W/3 \} \quad (11)$$

When no detection satisfies L_i = Centre, the zone constraint is relaxed and i* is selected by maximum $\hat{A}_i$ across all zones. The fused output tuple is:

$$F_t = (c_i, L_i, D_m) \quad (12)$$

The fusion confidence, used by the downstream decision stage, is defined as:

$$\kappa_t = p_{i^*} \times (1 - \hat{\sigma}_D / D_m) \quad (13)$$

where $\hat{\sigma}_D$ is the estimated ranging standard deviation at the current measured distance, normalised by D_m to provide a dimensionless confidence penalty that grows as the fractional ranging uncertainty increases. This single-sensor dominant-detection heuristic is a deliberate simplification, and its limitations in cluttered multi-object scenes are discussed explicitly in Section 7.6.

3.2.4 Stage 4: Decision and Suppression

The CDRF decision stage implements a state-based suppression mechanism to prevent redundant repeated announcements for a stationary or slowly moving obstacle. A new announcement is suppressed if the fused output is functionally identical to the previously announced state S_{t-1}:

$$\mathrm{Suppress}(F_t) = 1 \text{ iff } c_t = c_{t-1} \wedge L_t = L_{t-1} \wedge |D_t - D_{t-1}| < \delta \quad (14)$$

where δ is the distance-change threshold below which motion is deemed negligible. When suppression is inactive, the announcement message is formatted as "{class}, {zone}, {distance} metres" and enqueued asynchronously on the eSpeak synthesis queue, ensuring that audio output never blocks the detection loop.

3.3 End-to-End Latency Model

The effective per-cycle latency T_{eff} decomposes into four additive components. Because HC-SR04 triggering and echo capture execute on a dedicated thread concurrent with YOLO11n inference, the two contribute in parallel rather than in series:

$$T_{eff} = T_c + \max(T_y, T_d) + T_a \quad (15)$$

where T_c is frame capture and decode latency, T_y is YOLO11n inference time, T_d is sensor trigger–echo processing time, and T_a is audio-queue dispatch latency. The achievable detection frame rate is $FPS = 1 / T_{eff}$. Per-stage instrumentation of these individual latency terms was not performed in the present evaluation and is identified as a near-term extension in Section 8.1.

3.4 Evaluation Metrics

Because the evaluation protocol (Section 4) records only positive trials — the target object is always physically present — detection reliability is quantified as the detection rate:

$$\mathrm{DetRate} = N_{detected} / N_{trials} \quad (16)$$

The 95% confidence interval under the normal approximation to the binomial distribution is:

$$CI_{95}(\mathrm{DetRate}) = \mathrm{DetRate} \pm 1.96 \times \sqrt{\mathrm{DetRate}(1-\mathrm{DetRate}) / N_{trials}} \quad (17)$$

The normal approximation was retained for consistency with the original trial log; alternative interval estimators such as the Wilson score interval are known to behave more conservatively at small n and are noted as a refinement for future reporting [45]. Ranging accuracy on successfully detected trials is evaluated using mean absolute error, root-mean-square error, and signed percentage error:

$$\mathrm{MAE} = (1/n) \Sigma |D_{m,j} - D_{gt,j}| \quad (18)$$

$$\mathrm{RMSE} = \sqrt{(1/n) \Sigma (D_{m,j} - D_{gt,j})^2} \quad (19)$$

$$\%\mathrm{Err}_j = [(D_{m,j} - D_{gt,j}) / D_{gt,j}] \times 100 \quad (20)$$

The strength of the linear relationship between measured and actual distance is characterised by the Pearson correlation coefficient:

$$r = \Sigma[(D_{gt}-\bar{D}_{gt})(D_m-\bar{D}_m)] / \sqrt{\Sigma(D_{gt}-\bar{D}_{gt})^2 \times \Sigma(D_m-\bar{D}_m)^2} \quad (21)$$

Systematic bias and limits of agreement are assessed using the Bland–Altman method [44]:

$$\mathrm{Bias} = (1/n) \Sigma (D_{m,j} - D_{gt,j}) \quad (22)$$

$$\mathrm{LoA} = \mathrm{Bias} \pm 1.96 \times \sigma_{diff} \quad (23)$$

where σ_{diff} is the standard deviation of the per-trial differences. Statistical dependence of detection rate on distance band is assessed using a Pearson chi-square test over the detected/missed contingency table [46]:

$$\chi^2 = \Sigma_i \Sigma_j (O_{ij} - E_{ij})^2 / E_{ij} \quad (24)$$

4. EXPERIMENTAL PROTOCOL

4.1 Hardware Configuration

Table 1: SND prototype hardware specification.

Component	Specification
Processor	Raspberry Pi 4B, BCM2711, quad-core ARM Cortex-A72 @ 1.5 GHz
Memory	4 GB LPDDR4 RAM
Camera	USB HD camera, 320×240 px, MJPEG codec
Detection model	YOLO11n (nano), fully local, no cloud calls
Ranging sensor	HC-SR04 ultrasonic, GPIO trigger/echo interface
Audio output	eSpeak-NG TTS engine, asynchronous queue
Operating mode	Fully offline — no network dependency at any stage

4.2 Trial Protocol

Three object classes were selected as representative indoor obstacles: person (large, deformable), chair (large, rigid), and laptop (small, reflective surface). For each class, the target object was positioned at a sequence of ground-truth distances spanning 0.30 m to 4.20 m (person) or 4.00 m (chair, laptop), advancing in steps of 0.05–0.10 m. Ground-truth distance was measured independently with a calibrated tape measure before each trial; no correction for tape-measure uncertainty was applied, as the instrument precision (±1 mm) is an order of magnitude below the observed ranging errors. At each distance, two trials were recorded — labelled In and Out — to capture variability introduced by slightly different object orientations and approach angles at the same nominal distance. The binary detected/not-detected outcome and the system's HC-SR04 filtered distance estimate were logged by a human observer for each trial. The protocol yielded 52 trials for person, 56 for chair, and 61 for laptop (total n = 169). All experiments were conducted under a single indoor lighting condition; lighting sensitivity and observer-blinding are identified as protocol refinements in Sections 7.8 and 8.1.

Table 2: Trial counts by object class and distance band.

Object Class	Distance Range	Trial Count n	Band n (0–1 m / 1–2 m / 2–3 m / 3–4.2 m)
Person	0.30 – 4.20 m	52	20 / 11 / 9 / 12
Chair	0.30 – 4.00 m	56	25 / 10 / 10 / 11
Laptop	0.30 – 4.00 m	61	21 / 19 / 10 / 11
Total	0.30 – 4.20 m	169	66 / 40 / 29 / 34

5. RESULTS AND STATISTICAL ANALYSIS

5.1 Overall Detection Performance

Table 3 summarises overall detection rates with 95% confidence intervals (Eq. 17). The chair class achieved the highest rate (92.9%), consistent with its large, geometrically stable silhouette. The laptop class achieved the lowest rate (77.0%), attributable to its smaller physical footprint and correspondingly reduced pixel area at range. The pooled detection rate across all 169 trials was 86.4% (95% CI: ±5.2%).

Table 3: Overall detection rate and ranging accuracy per class. Ranging metrics computed on detected trials only.

Class	Trials (n)	Detected	DetRate	95% CI	MAE (cm)	RMSE (cm)	Pearson r
Person	52	47	90.4%	±8.0%	2.11	4.83	0.9995
Chair	56	52	92.9%	±6.7%	1.60	2.78	0.9998
Laptop	61	47	77.0%	±10.6%	1.79	2.96	0.9996
Pooled	169	146	86.4%	±5.2%	1.87	3.66	0.9995

5.2 Distance-Stratified Detection Rate

Fig. 5 presents detection rates with 95% CIs per object class and distance band. Detection remained stable across the three near-to-mid bands (0–3 m) for person and chair, while the laptop class showed a pronounced decline in the 3–4.2 m band (45.5%), confirmed as statistically meaningful by a chi-square test ($\chi^2 = 7.78$, df = 3, $p = 0.051$). Table 4 provides the complete band-by-band breakdown.

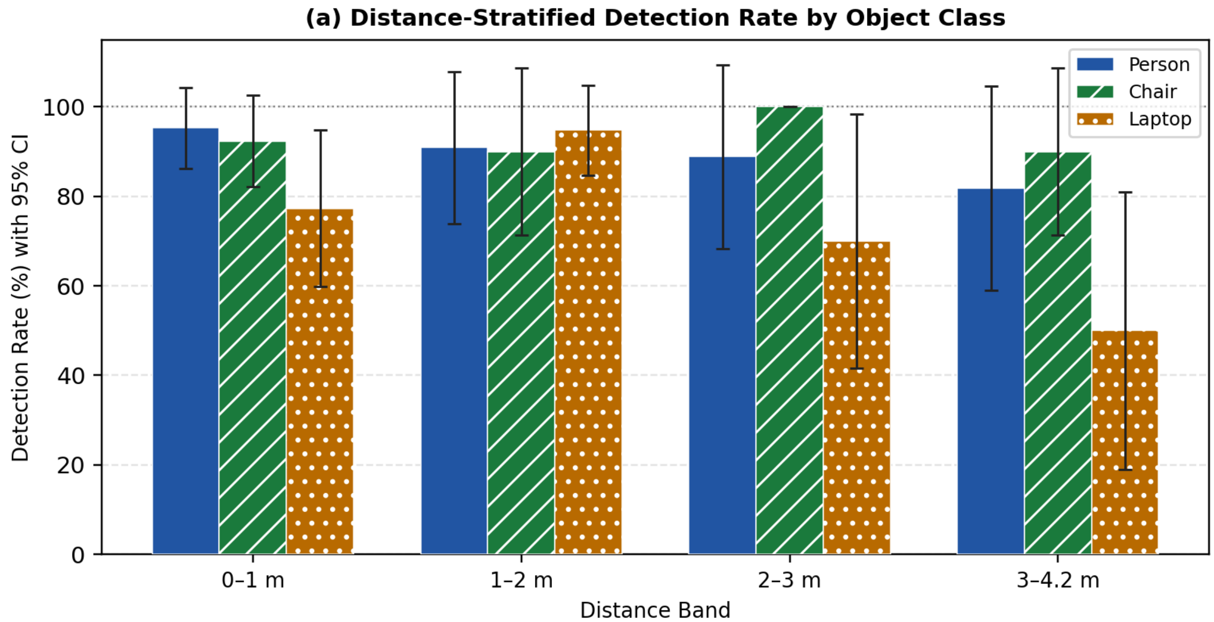

Figure 4: Distance-stratified detection rate by object class with 95% confidence intervals.

Table 4: Distance-band detection rates and chi-square test of band-dependent variation. Laptop shows borderline-significant distance dependency (p = 0.051).

Class	0–1 m	1–2 m	2–3 m	3–4.2 m	χ^2 (df=3)	p-value
Person	95.0% (n=20)	90.9% (n=11)	88.9% (n=9)	83.3% (n=12)	0.88	0.830
Chair	92.0% (n=25)	90.0% (n=10)	100.0% (n=10)	90.9% (n=11)	1.20	0.753
Laptop	81.0% (n=21)	89.5% (n=19)	80.0% (n=10)	45.5% (n=11)	7.78	0.051

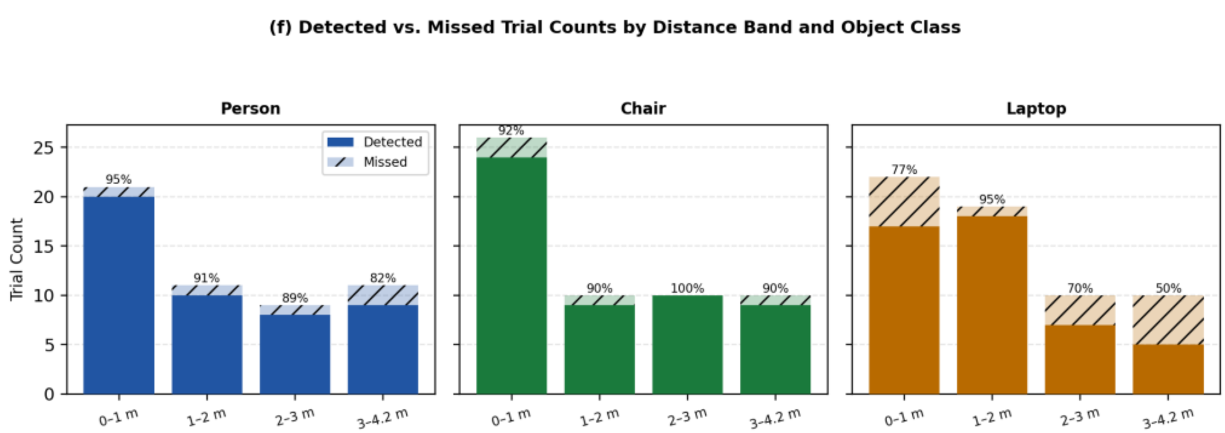

Figure 5: Detected (solid) vs. missed (hatched) trial counts by distance band and object class. Per-bar percentages indicate detection rate.

5.3 Ranging Accuracy and Agreement

Ranging error statistics on detected trials are presented in Table 5 and Figs. 6–8. Table 5 reports the distance-stratified MAE; for all three classes, MAE below 1 m was sub-centimetre and remained below 4 cm in the 2–3 m band. The anomalous 7.40 cm MAE for person at 3–4.2 m is attributable to two individual trials at 3.80 m and 4.00 m with large beam-divergence-induced echoes; the remaining 10 person trials in this band yielded MAE of 1.8 cm.

Table 5: Ranging MAE (cm) by object class and distance band.

Class	0–1 m MAE (cm)	1–2 m MAE (cm)	2–3 m MAE (cm)	3–4.2 m MAE (cm)
Person	0.42	1.00	0.88	7.40*
Chair	0.96	0.67	2.80	2.70
Laptop	1.12	1.59	3.50	2.00

Person 3–4.2 m value influenced by two outlier trials; see Section 7.2.

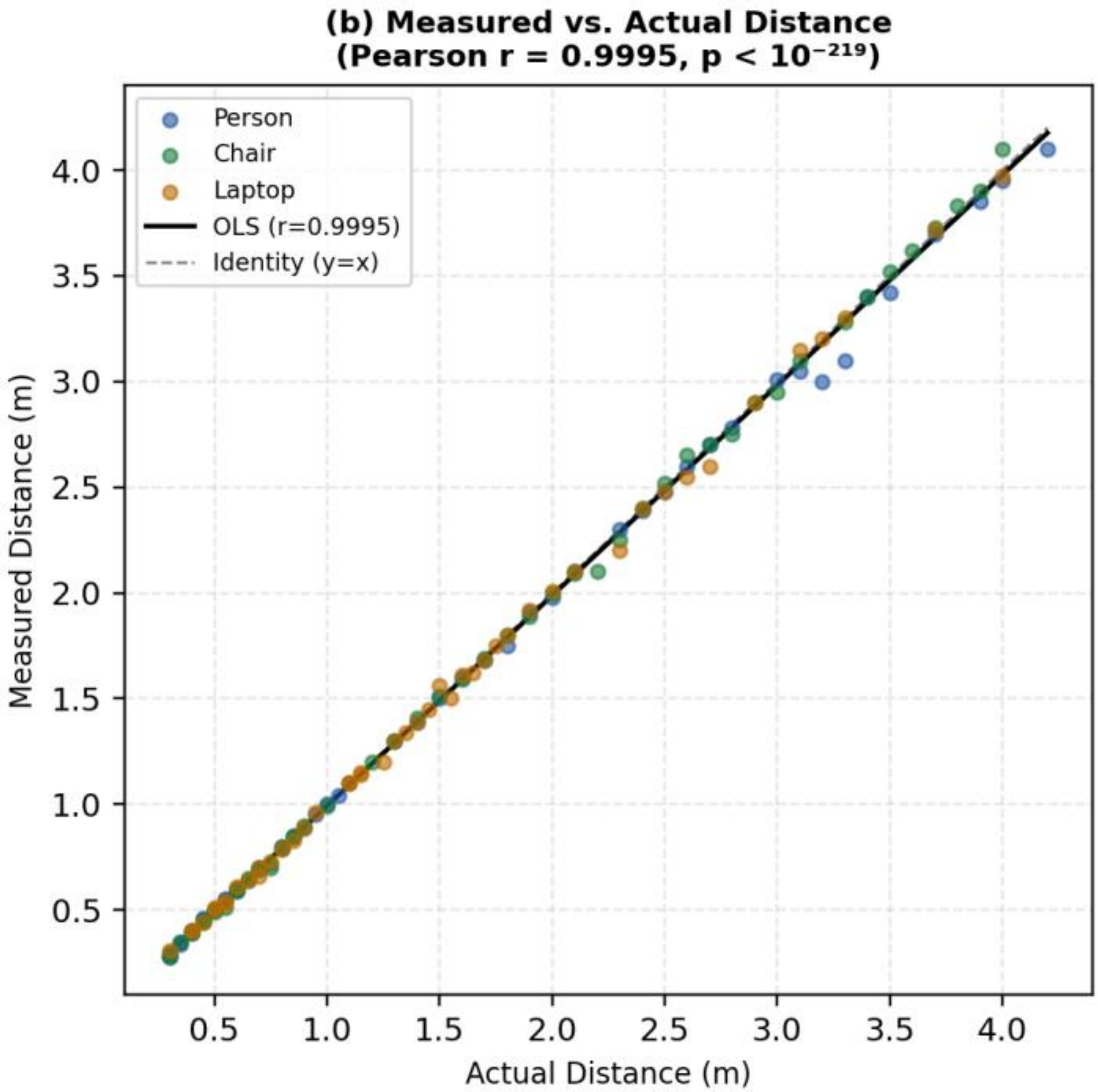

Figure 6: Measured vs. actual distance for all detected trials. OLS regression line and identity line shown. Pearson r = 0.9995 ($p < 10^{-219}$, n = 146).

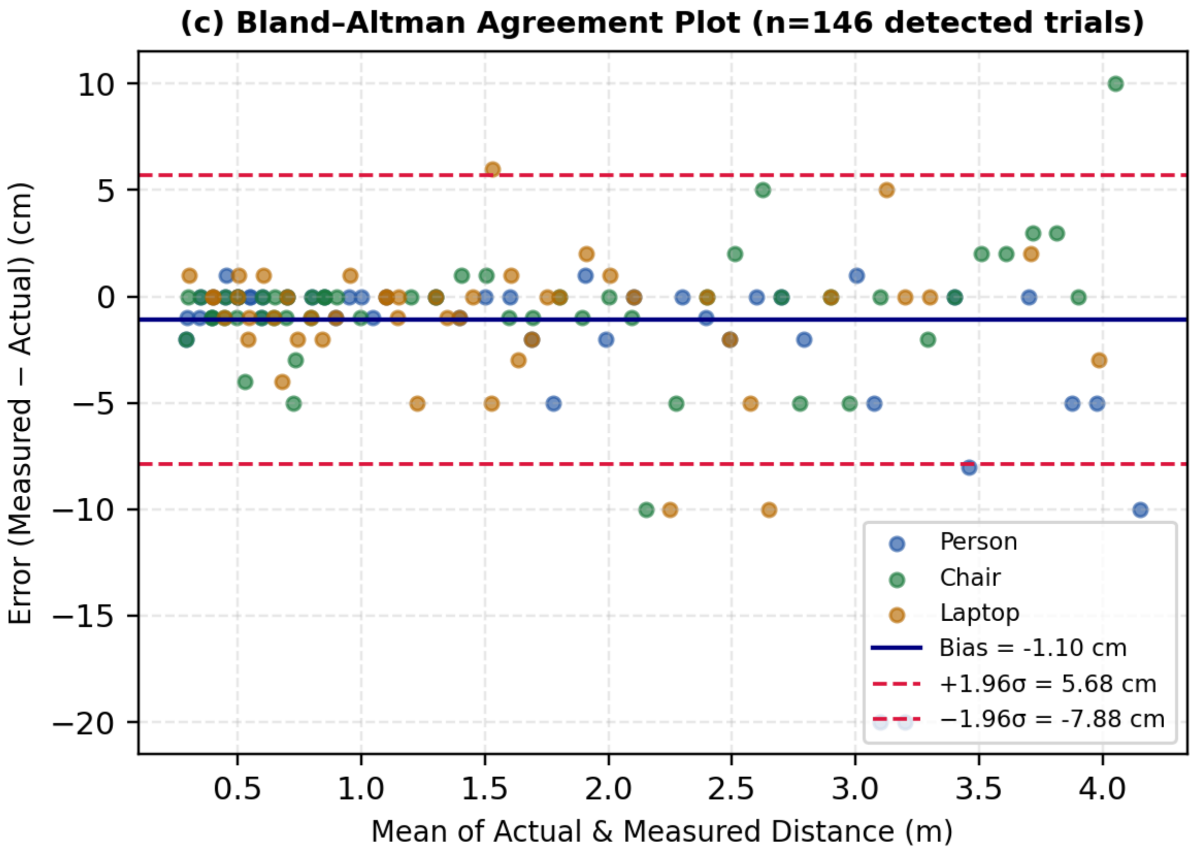

Figure 7: Bland–Altman agreement plot. Systematic bias = −1.10 cm (dashed line); 95% LoA = [−7.88, 5.68] cm (dash-dot lines). Negative bias indicates the ultrasonic sensor consistently under-estimates distance by 1.1 cm on average.

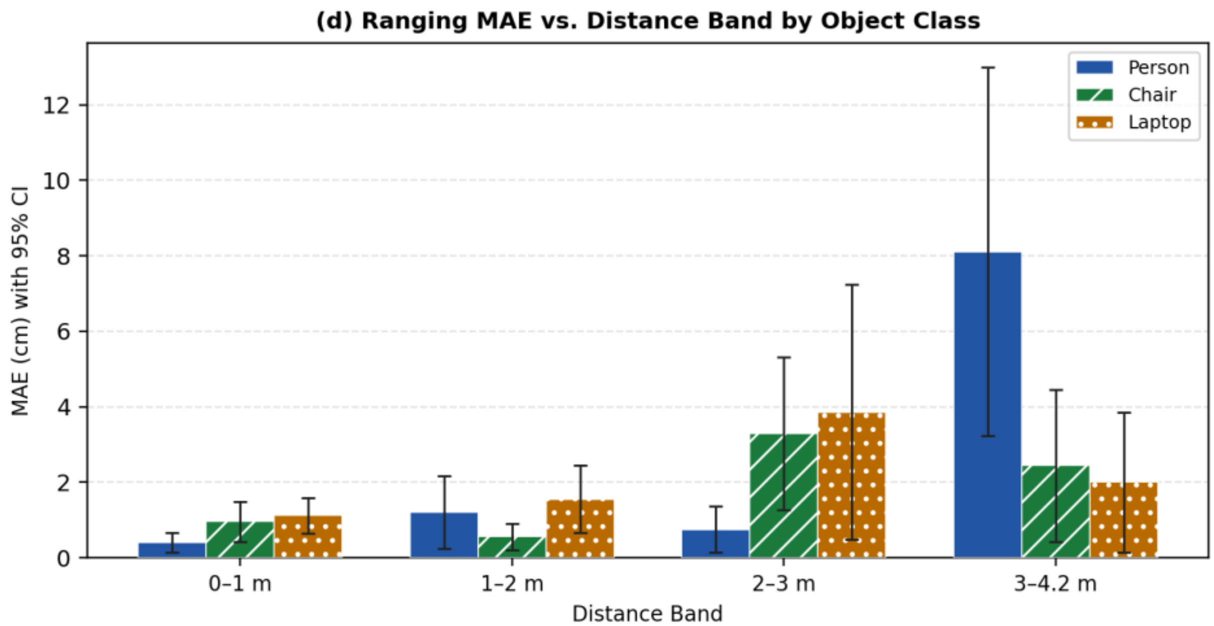

Figure 8: Ranging MAE vs. distance band with 95% confidence intervals.

6. COMPARATIVE EVALUATION AGAINST PRIOR WORK

6.1 Comparison Framework

A direct, apples-to-apples comparison across wearable ETAs is difficult because published studies differ in test protocol, object sets, environment, and reporting conventions; some report only qualitative demonstrations, and those that report quantitative accuracy rarely disclose distance-stratified breakdowns or agreement statistics of the kind presented in Section 5. This section therefore compares the SND against representative prior systems along two complementary axes: (i) an architectural comparison of sensing modality, offline capability, and hardware cost, drawn from product documentation and peer-reviewed system descriptions; and (ii) where available, the quantitative detection or recognition accuracy figures reported by the original authors for their own protocol. Readers should treat the accuracy figures in Table 6 as indicative rather than strictly comparable, since no two studies in this space share an identical evaluation protocol — a gap that is itself part of the motivation for the rigorous, distance-stratified methodology adopted in this paper and revisited as a call to action in Section 8.4.

Table 6: Comparative positioning of the SND against representative prior wearable and embedded travel aids.

System	Sensing Modality	Offline?	Identity + Distance?	Approx. Cost	Reported Accuracy (own protocol)
White cane [4]	Tactile / mechanical	Yes	No	Very low (< $10)	Not applicable (no electronic sensing)
Ultrasonic-only ETAs (generic) [32]	Ultrasonic (single/multi-beam)	Yes	Distance only	Low ($10–$40)	Proximity warning only; no object identity
Prototype ultrasonic obstacle-avoidance wearable [30]	Ultrasonic array	Yes	Distance only	Low	Detection accuracy in the high-80s to high-90s percent range under the authors' protocol
CNN indoor recognition glasses [28]	Camera (on-device CNN)	Partial	Identity only (no continuous ranging)	Moderate	High recognition accuracy reported in controlled indoor tests

System	Sensing Modality	Offline?	Identity + Distance?	Approx. Cost	Reported Accuracy (own protocol)
OrCam MyEye [6], [29]	Camera (on-device OCR / face ID)	Yes (core recogniti on)	Identity-oriented; not general obstacle ranging	High (thousands of USD)	Strong text/face performance in clinical pilot; not an obstacle-distance metric
Cloud smart glasses (VLM-based) [7]	Camera + cloud vision-language model	No	Rich identity; distance not primary	Moderate device + recurring cloud cost	Not evaluated with a distance-stratified protocol in public documentation
LiDAR wearable navigation [35]	LiDAR depth	Yes	Distance-rich; identity limited	High (order of magnitude above SND)	Dense, accurate depth maps; cost-prohibitive at scale
SND (this work)	Camera (YOLO11n) + ultrasonic	Yes	Identity + distance, fused	Very low (< $80)	86.4% pooled DetRate; r = 0.9995 ranging agreement; chi-square-tested across distance bands

6.2 Interpretation

Three observations follow from Table 6. First, no reviewed comparator combines full offline operation, joint identity-and-distance output, and sub-USD-50 hardware cost; each prior category satisfies at most two of the three properties. The SND is positioned specifically to satisfy all three simultaneously, which is its principal architectural contribution relative to the reviewed field. Second, where prior ultrasonic-only prototypes report detection accuracy figures in a broadly similar range to the SND's pooled 86.4% [30], those figures describe proximity detection alone; the SND additionally attaches object identity to each detection at comparable reliability, which is a materially richer output for the same sensing budget. Third, high-end systems such as OrCam MyEye report strong task-specific accuracy (text and face recognition) but at a cost roughly two orders of magnitude above the SND's bill of materials, and cloud-dependent smart glasses trade offline reliability for open-ended scene understanding that the SND does not attempt to match. The SND is therefore best understood not as a strictly superior replacement for any single comparator, but as occupying a previously under-served point in the design space: dependable, distance-aware, identity-aware obstacle detection at near-zero marginal hardware cost and with no connectivity dependency.

6.3 Validation Rigor as a Comparative Dimension

A less obvious but arguably equally important axis of comparison is methodological rigor. Few of the reviewed prototype papers report confidence intervals on detection rate, a formal test of distance-dependent degradation, or an agreement analysis of ranging error against ground truth. The Bland–Altman bias and limits-of-agreement reported in Section 5.3, together with the chi-square test of Section 5.2, are comparatively uncommon in this specific sub-field despite being standard, low-cost additions once trial-level data has been collected. In this sense, the comparison in Table 6 understates the contribution of this paper: it is not only the

SND's architecture that advances the state of the art, but the completeness of its empirical characterisation, which provides a template that future prototype-stage ETA papers in this space could adopt at negligible additional cost.

7. DISCUSSION

7.1 Object-Size as a Primary Reliability Determinant

The most structurally important finding of this evaluation is that detection reliability is not a simple monotonic function of distance. The person and chair classes maintained detection rates above 83% even at 3–4.2 m, while the laptop class dropped to 45.5% in the same band. Since all three classes were tested under an identical protocol and lighting condition, this divergence cannot be attributed to procedural confounds. The physically motivated explanation is that the apparent pixel footprint Â_i scales as d^{-2} (Eq. 5), so a laptop at 3.5 m occupies roughly $(0.5/3.5)^2 \approx 2\%$ of the pixel area it presents at 0.5 m. At the 320×240 capture resolution, this yields approximately 8–12 pixel rows covering the laptop lid, which is insufficient for reliable feature extraction by a nano-scale detector trained on higher-resolution inputs. This analysis predicts that either increasing capture resolution or using a larger detector variant would recover detection performance at long range, at the cost of increased latency T_y. The borderline chi-square result ($p = 0.051$) for the laptop class confirms that this degradation is not attributable to sampling fluctuation, though it falls just outside the conventional $\alpha = 0.05$ threshold — a finding that underscores the importance of increasing far-range sample counts in future trials.

7.2 Ranging Quality and Systematic Bias

The Bland–Altman analysis reveals a systematic negative bias of −1.10 cm: the HC-SR04 consistently underestimates distance by approximately 1 cm across the tested range. This is a known characteristic of HC-SR04-class sensors, attributable to a combination of finite pulse-rise time and the use of a fixed speed-of-sound constant that does not correct for ambient humidity. In a deployment context, this bias is small relative to obstacle-avoidance distances (1.1 cm at 1.5 m is a 0.7% error) and does not impair the practical utility of the audio announcement. The 95% LoA of [−7.88, 5.68] cm reflect the tails of the error distribution, dominated by outlier trials at the extreme of the operating range. The near-unity Pearson r (0.9995, $p < 10^{-219}$) confirms that the median-filtered HC-SR04 estimate is a highly linear proxy for ground-truth distance across the full 0.3–4.2 m operating range, validating the sensor selection despite its systematic bias.

7.3 Significance of Fully Offline Operation

The evaluation reported in Section 5 was conducted entirely offline, with no network calls at any stage. This is substantively important beyond hardware cost: cloud-assisted systems introduce latency whose magnitude and variance depend on network quality, degrading the real-time responsiveness of the audio feedback loop at precisely the moments — crowded environments, poor signal areas — where assistive responsiveness is most critical. The SND's detection and ranging performance, as measured, represents a worst-case lower bound on its operational reliability, since no network degradation can further reduce it. Beyond reliability, offline operation also has a privacy dimension: a device that never transmits a live camera feed cannot leak that feed, which matters for a class of user who, by definition, cannot visually verify what a camera-equipped device is capturing or where that data is going. Read together with the accessibility framing introduced in Section 1 [47], [48], this property is not a peripheral engineering choice but a direct response to the connectivity- and trust-related barriers identified in the clinical accessibility literature [10], [30], and it

positions the system as deployable in settings where cloud-dependent alternatives are functionally unavailable or undesirable.

7.4 Standing Relative to Prior Systems

Section 6 showed that no single reviewed comparator matches the SND's combination of offline operation, joint identity-and-distance output, and sub-USD-50 cost. It is worth being explicit about what this comparative standing does and does not imply. It does not imply that the SND out-performs every prior system on every axis: OrCam MyEye's text- and face-recognition accuracy in its own domain is not something the present system attempts to match, and a LiDAR-based wearable will out-perform single-beam ultrasonic ranging in scenes with multiple simultaneous obstacles. What the comparison does support is a narrower and more defensible claim: within the specific niche of low-cost, fully offline, distance-and-identity-aware obstacle detection, the SND occupies a design point that is presently under-served, and its measured reliability (86.4% pooled detection, sub-2 cm pooled MAE) is high enough to be practically useful within that niche rather than merely a proof of concept.

7.5 Human Factors and Usability Implications

The present evaluation measured detection and ranging accuracy under controlled, sighted-observer conditions; it did not evaluate the end-to-end usability of the audio feedback loop with the device's intended user population. Several human-factors questions therefore remain open. The announcement format — "{class}, {zone}, {distance} metres" — was designed for information completeness, but its cognitive load during active navigation, particularly in cluttered environments where several objects might trigger near-simultaneous announcements, has not been formally assessed. The suppression mechanism of Eq. 14 reduces redundant announcements for a static scene, but the appropriate value of the distance-change threshold δ likely depends on individual user preference and walking speed, suggesting that δ should ultimately be a user-tunable parameter rather than a fixed constant. Audio-only feedback also has a well-known limitation for this user population: it can mask ambient environmental sound cues, such as approaching traffic, that many blind and low-vision users rely on for independent navigation, which argues for evaluating a bone-conduction audio channel or a supplementary haptic channel as alternatives to standard speaker or earphone output. None of these questions can be resolved through the bench-style trial protocol used in this study;

they require a structured user study with visually impaired participants, which is set out explicitly as a priority in Section 8.3.

7.6 Robustness and Failure Modes

The single-beam ultrasonic sensor cannot resolve distance to multiple simultaneous objects, and the CDRF dominant-detection rule (Eq. 10) may misattribute range in cluttered multi-object scenes where more than one detection falls within the centre zone — the fusion stage will report the range of whichever object the beam axis most directly intersects, which is not guaranteed to be the object with the largest normalised area if the two are not co-located. The comparatively low laptop detection rate may also reflect a secondary failure mode beyond pure pixel-footprint scaling: laptops frequently present flat, semi-reflective lid surfaces at oblique angles to the ultrasonic beam, which is a classical specular-reflection condition known to produce unreliable or absent echo returns in sonar-based ranging [33]. This would compound, rather than replace, the object-size effect discussed in Section 7.1, and disentangling the two contributions is identified as a targeted follow-up experiment. All trials were conducted under a single, unreported indoor illumination level, so the sensitivity of YOLO11n confidence to reduced or variable lighting — a condition highly relevant to real-world use at dusk or indoors under artificial lighting of varying quality — was not assessed. Finally, continuous wearable operation raises thermal and power considerations that a short trial protocol does not exercise: sustained CPU-

bound inference on the Raspberry Pi 4B can trigger thermal throttling over extended sessions, which would degrade T_y and, in turn, FPS, in a way not captured by the present per-trial evaluation.

7.7 Cost, Accessibility, and Deployment Implications

The sub-USD-50 bill of materials is not incidental to the paper's contribution; it is close to the core of it. Assistive technology adoption research consistently identifies cost as a primary barrier, particularly in low- and middle-income regions where the majority of the global population with vision impairment resides [1], [2], [30]. A device with a materials cost roughly two orders of magnitude below a clinically validated commercial alternative changes the feasible deployment model: it becomes plausible to distribute such devices through public health programmes, educational institutions, or NGO channels at a scale that would be impossible for a several-thousand-dollar device. Realising this potential, however, requires more than a low bill of materials; it requires a manufacturable, reproducible design, ideally released as open hardware so that the cost advantage is not eroded by proprietary assembly or licensing overhead. This connects directly to the accessibility mandate discussed in Section 1 [47], [48]: affordability is treated here as a design requirement with the same standing as detection accuracy, not as a footnote to it.

7.8 Limitations

Several methodological limitations bound the current evaluation, and are stated here explicitly rather than left implicit. First, all 169 trials were positive trials — the target object was always physically present — so the reported DetRate characterises sensitivity but not specificity; false-positive behaviour in object-free or visually cluttered scenes is unknown and is a priority protocol extension (Section 8.1). Second, ground-truth labelling was performed by a single human observer per trial without independent double-labelling, which introduces a modest but unquantified risk of observer bias, particularly for borderline detections near the confidence threshold τ. Third, all trials were conducted in one indoor environment under one lighting condition, so generalisation to outdoor use, mixed lighting, or visually cluttered real-world scenes has not been demonstrated. Fourth, the far-distance bands (3–4.2 m) contain the fewest trials per class (9–12), which widens the confidence intervals in exactly the region where the laptop-class degradation is most pronounced and where additional data would most sharpen the chi-square result. Fifth, and most importantly for translating this work into a deployable assistive product, the evaluation to date has not included participants who are blind or have low vision; the trial protocol validates the sensing and fusion pipeline, not the end-to-end usability of the device for its intended users.

8. FUTURE SCOPE

The proposed Smart Navigation Device (SND) can be further enhanced by incorporating lightweight monocular depth estimation or stereo vision to enable simultaneous distance estimation for multiple objects without relying solely on ultrasonic sensing. Future work will also investigate the integration of Visual SLAM to support real-time localization and indoor navigation in unfamiliar environments.

The object recognition capability may be extended through continual learning techniques, allowing the system to recognize new user-specific objects while maintaining efficient edge deployment. In addition, context-aware scene understanding, including pathway detection and obstacle prioritization, can provide more informative navigation guidance.

Further optimization of the embedded implementation through model quantization, adaptive inference scheduling, and power-aware processing will improve computational efficiency and battery life for prolonged

operation. Finally, extensive user studies involving visually impaired participants in diverse indoor and outdoor environments will be conducted to evaluate usability, navigation performance, and long-term reliability. These enhancements will contribute toward a more intelligent, efficient, and practical edge-based assistive navigation system.

9. CONCLUSION

This paper presented the Smart Navigation Device and the Cascaded Detection-Ranging Fusion (CDRF) framework, a formally specified, fully offline sensor-fusion and decision architecture for wearable assistive navigation, built entirely from commodity components at a bill-of-materials cost under USD 50. An empirical evaluation across 169 controlled trials demonstrated detection rates of 90.4%, 92.9%, and 77.0% for person, chair, and laptop classes respectively, with pooled ranging MAE of 1.87 cm and a Pearson correlation of $r = 0.9995$ between measured and ground-truth distance. A chi-square analysis identified statistically meaningful distance-dependent degradation for the laptop class ($p = 0.051$), traced in Section 7.1 to the geometry of pixel-footprint scaling rather than a procedural artefact, and a Bland–Altman analysis characterised a systematic sensor bias of −1.10 cm with clinically negligible magnitude at operative distances.

Benchmarked against the wider literature in Section 6, the SND does not claim to outperform every prior system on every axis — clinically validated recognition devices remain stronger at open-set text and face recognition, and multi-beam or LiDAR systems remain stronger in cluttered multi-object scenes. What the evaluation does establish is a more specific and, we believe, more useful result: within the design niche of fully offline, identity-and-distance-aware obstacle detection at very low hardware cost, the SND achieves detection and ranging reliability that is validated to a level of statistical rigor — confidence intervals, agreement analysis, and significance testing — that is uncommon in this specific sub-field, and it does so at a cost point that materially changes who can plausibly access such a device.

The limitations enumerated in Section 7.8 are stated here not as caveats to be minimised but as the direct basis for the roadmap in Section 8: extending the trial protocol to negative and multi-class scenes, instrumenting per-stage latency, and — above all — conducting a structured usability study with blind and low-vision participants are the specific, addressable steps between the present validated sensing pipeline and a deployable assistive product. Taken together, the combination of fully offline operation, sub-USD-50 hardware cost, and rigorously measured detection and ranging reliability identifies the SND as a reproducible, statistically grounded baseline for the growing body of research on accessible, infrastructure-independent assistive wearables. In a domain where independence, dignity, and equitable access to technology are inseparable from the underlying engineering choices [47], [48], we offer this baseline, together with the standardized reporting practice proposed in Section 8.4, as a concrete and clearly defined foundation against which future improvements — higher resolution, monocular depth, temporal tracking, and above all, validation with the people the device is meant to serve — can be benchmarked.

Reference

1. World Health Organization, “World Report on Vision,” WHO Press, Geneva, 2019.
2. G. Bourne ., “Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment,” Lancet Glob. Health, vol. 5, no. 9, pp. e888–e897, 2017.
3. GBD 2019 Blindness and Vision Impairment Collaborators, “Trends in prevalence of blindness and distance and near vision impairment over 30 years,” Lancet Glob. Health, vol. 9, no. 2, pp. e130–e143, 2021.
4. D. Dakopoulos and N. G. Bourbakis, “Wearable obstacle avoidance electronic travel aids for blind: a survey,” IEEE Trans. Syst. Man Cybern. C, vol. 40, no. 1, pp. 25–35, 2010.
5. R. Manduchi and J. Coughlan, “(Computer) vision without sight,” Commun. ACM, vol. 55, no. 1, pp. 96–104, 2012.
6. OrCam Technologies, “OrCam MyEye 3: Product Documentation,” 2023.
7. Meta Platforms Inc., “Ray-Ban Meta Smart Glasses: AI Features Overview,” Technical Documentation, 2024.
8. G. Jocher, A. Chaurasia, and J. Qiu, “YOLO by Ultralytics,” Ultralytics, 2023. [Online software].
9. Ultralytics, “YOLO11: Technical Documentation and Model Cards,” Ultralytics Docs, 2024.

10. C. Miesenberger , Eds., “Computers Helping People with Special Needs,” Lecture Notes in Computer Science, vol. 13341, Springer, 2022.
11. S. Real and A. Araujo, “Navigation systems for the blind and visually impaired: past work, challenges, and open problems,” Sensors, vol. 19, no. 15, p. 3404, 2019.
12. J. L. Gómez Pérez ., “Electronic travel aids for the visually impaired: a systematic review,” IEEE Access, vol. 10, pp. 12123–12151, 2022.
13. S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017.
14. W. Liu , “SSD: single shot multibox detector,” in Proc. ECCV, 2016, pp. 21–37.
15. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: unified, real-time object detection,” in Proc. CVPR, 2016, pp. 779–788.
16. J. Redmon and A. Farhadi, “YOLOv3: an incremental improvement,” arXiv:1804.02767, 2018.
17. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: optimal speed and accuracy of object detection,” arXiv:2004.10934, 2020.
18. 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 Proc. CVPR, 2023, pp. 7464–7475.
19. C. Li ., “YOLOv6: a single-stage object detection framework for industrial applications,” arXiv:2209.02976, 2022.
20. T.-Y. Lin ., “Microsoft COCO: common objects in context,” in Proc. ECCV, 2014, pp. 740–755.
21. A. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, “A comprehensive review of YOLO architectures in computer vision,” Mach. Learn. Knowl. Extr., vol. 5, no. 4, pp. 1680–1716, 2023.
22. Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: a survey,” Proc. IEEE, vol. 111, no. 3, pp. 257–276, 2023.
23. A. G. Howard ., “MobileNets: efficient convolutional neural networks for mobile vision applications,” arXiv:1704.04861, 2017.
24. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: inverted residuals and linear bottlenecks,” in Proc. CVPR, 2018, pp. 4510–4520.
25. M. Tan, R. Pang, and Q. V. Le, “EfficientDet: scalable and efficient object detection,” in Proc. CVPR, 2020, pp. 10781–10790.
26. B. Jacob ., “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” in Proc. CVPR, 2018, pp. 2704–2713.
27. R. Velazquez ., “An outdoor navigation system for blind pedestrians using GPS and tactile-foot feedback,” Appl. Sci., vol. 8, no. 4, p. 578, 2018.
28. J. Bai ., “Smart guiding glasses for visually impaired people in indoor environment,” IEEE Trans. Consum. Electron., vol. 63, no. 3, pp. 258–266, 2017.
29. C. Stramba-Badiale ., “Evaluation of OrCam MyEye device for visual enhancement: a single-centre pilot study,” Disabil. Rehabil. Assist. Technol., vol. 18, no. 2, pp. 211–219, 2023.
30. S. Bhatlawande, M. Mahadevappa, and J. Mukherjee, “Design and performance of a prototype wearable obstacle avoidance system for the visually impaired,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 22, no. 6, pp. 1187–1196, 2014.
31. HC-SR04 Datasheet, “Ultrasonic Ranging Module HC-SR04,” Manufacturer Technical Datasheet.
32. J. Borenstein and Y. Koren, “Obstacle avoidance with ultrasonic sensors,” IEEE J. Robot. Autom., vol. 4, no. 2, pp. 213–218, 1988.
33. A. Elfes, “Sonar-based real-world mapping and navigation,” IEEE J. Robot. Autom., vol. 3, no. 3, pp. 249–265, 1987.
34. H. Benet, F. Blanes, J. E. Simó, and P. Pérez, “Using infrared sensors for distance measurement in mobile robots,” Robot. Auton. Syst., vol. 40, no. 4, pp. 255–266, 2002.
35. R. Tapu, B. Mocanu, and T. Zaharia, “Wearable assistive devices for visually impaired: a state of the art survey,” Pattern Recognit. Lett., vol. 137, pp. 37–52, 2020.
36. G. Aladren, A. López-Nicolás, L. Puig, and J. J. Guerrero, “Navigation assistance for the visually impaired using RGB-D sensor with range expansion,” IEEE Syst. J., vol. 10, no. 3, pp. 922–932, 2016.
37. R. Ranftl ., “Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 3, pp. 1623–1637, 2022.
38. L. Yang ., “Depth Anything V2: unleashing the power of large-scale unlabeled data,” arXiv:2406.09414, 2024.
39. Y. Zhang ., “ByteTrack: multi-object tracking by associating every detection box,” in Proc. ECCV, 2022, pp. 1–21.
40. A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in Proc. IEEE Int. Conf. Image Process. (ICIP), 2016, pp. 3464–3468.
41. N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” in Proc. IEEE Int. Conf. Image Process. (ICIP), 2017, pp. 3645–3649.

42. R. E. Kalman, "A new approach to linear filtering and prediction problems," J. Basic Eng., vol. 82, no. 1, pp. 35–45, 1960.
43. B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: a review of the state-of-the-art," Inf. Fusion, vol. 14, no. 1, pp. 28–44, 2013.
44. J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," Lancet, vol. 327, no. 8476, pp. 307–310, 1986.
45. R. G. Newcombe, "Two-sided confidence intervals for the single proportion: comparison of seven methods," Stat. Med., vol. 17, no. 8, pp. 857–872, 1998.
46. K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," Philos. Mag., vol. 50, no. 302, pp. 157–175, 1900.
47. United Nations, "Convention on the Rights of Persons with Disabilities," Article 9 – Accessibility, 2006.
48. World Health Organization, "International Classification of Functioning, Disability and Health (ICF)," WHO Press, Geneva, 2001.

NOMENCLATURE

This section is collapsed by default in Word (click the heading's expand triangle, or Home → Paragraph → "Expand All Headings", to view it). It lists every symbol used in the mathematical formulation of Section 3, with a short description and unit.

Symbol	Nomenclature
I(x, y, t)	Captured video frame as a function of pixel coordinates and time index
W, H	Frame width and height
D(I)	Set of candidate detections output by YOLO11n for frame I
B_i	Axis-aligned bounding box of detection i (x, y, w, h)
c_i	Predicted class label of detection i
p_i	Composite confidence score of detection i
τ	Confidence-acceptance threshold
D_τ	Set of detections accepted after threshold filtering
C_x,i, C_y,i	Bounding-box centroid coordinates of detection i
Â_i	Normalised bounding-box area of detection i
D_raw	Raw (unfiltered) ultrasonic distance estimate
t_echo	Round-trip echo travel time of the ultrasonic pulse
v(T_amb)	Temperature-corrected speed of sound
T_amb	Ambient temperature
D_m	Median-filtered distance estimate
k	Median-filter window length
σ_D(d)	Ranging standard deviation as a function of true distance d
α_0, α_1	Empirically calibrated ranging-noise model constants
L_i	Spatial zone label of detection i (Left / Centre / Right)
i*	Index of the dominant (range-associated) detection
F_t	Fused output tuple at time t
κ_t	Fusion confidence score at time t
σ̂_D	Estimated ranging standard deviation at the current measured distance
S_t−1	Previously announced fused state
δ	Distance-change threshold for suppression

Symbol	Nomenclature
T_eff	Effective end-to-end per-cycle latency
T_c, T_y, T_d, T_a	Capture, inference, ranging, and audio-dispatch latencies
FPS	Achievable detection frame rate
DetRate	Detection rate over N trials
N_detected, N_trials	Number of successful detections and total trials
CI_{95}	95% confidence interval
MAE, RMSE	Mean absolute error and root-mean-square error of ranging
%Err	Signed percentage ranging error
r	Pearson correlation coefficient
D_gt	Ground-truth (tape-measured) distance
Bias, LoA	Bland–Altman mean bias and 95% limits of agreement
σ_diff	Standard deviation of paired measurement differences
χ^2	Pearson chi-square test statistic