

GeoVision Cap: A Technology Gifted Eye – A Review of Geo-Localization and Wearable Assistive Technologies for The Visually Impaired

Nidhi Y. Pandya¹, Ritesh B. Vaghasiya²

¹*Department of Electronics and Communication, Dr. Subhash University, Junagadh India*

²*Assistant Professor., Department of Electronics and Communication, Dr Subhash University, Junagadh India*

Abstract—Visually impaired people rely on white canes, guide dogs and, increasingly, wearable electronics for safe mobility, yet most low-cost devices only warn of nearby obstacles and say little about where the user is or what surrounds them. This paper reviews five recent studies: three on visual geo-localization (global image geolocation, unmanned aerial vehicle (UAV) cross-view matching, and Mamba-based deep hashing) and two on cap-based assistive devices for blind users. The studies are compared by task, platform, method, dataset, reported results and suitability for wearable use. The review shows that geo-localization methods achieve strong benchmark accuracy but are developed for aircraft, vehicles or archived imagery on graphics processing unit (GPU) hardware, whereas assistive caps are inexpensive but limited to proximity warnings or slow object labelling, with little quantitative evaluation. Seven research gaps are identified. The design of GeoVision Cap, a Raspberry Pi 5 based edge artificial intelligence (Edge AI) wearable that combines vision, location and voice interaction, is then presented as a planned response to these gaps.

Index Terms—Assistive technology, Edge AI, Geo-localization, Raspberry Pi, Sensor fusion, Vision-language models, visually impaired, Wearable devices.

I. INTRODUCTION

Visual impairment limits independent mobility, object recognition and access to printed information. The traditional aids are the white cane and the guide dog. A white cane has difficulty detecting obstacles above knee level, such as street barriers [4], and guide dogs have limits of their own [5]. Smartphone applications and smart glasses add information, but smart glasses

remain expensive, which motivated the low-cost cap design reported in [5].

A cap or hat is an attractive wearable form factor because a camera and sensors mounted at head height face the walking direction and leave the hands free. Two recent studies build on this idea: a smart hat that detects objects with a deep learning model and speaks to the user through a smartphone application [4], and an ultrasonic cap that sounds a buzzer when an obstacle is near [5].

In parallel, computer vision research on geo-localization estimates where an image was taken, either without location metadata [1] or by matching it against satellite or remote-sensing imagery [2], [3]. These methods support navigation, robotics and positioning when satellite signals are unavailable, but they have not been connected to assistive wearables for visually impaired users.

This paper therefore reviews five studies spanning both directions, compares them in a common table, and identifies research gaps. It then describes the planned design of GeoVision Cap and maps each gap to a design response. GeoVision Cap is presented as a proposed design; this paper does not report experimental results for it.

II. REVIEW METHODOLOGY

A structured review of five studies was conducted. Three studies address visual geo-localization [1]– [3] and two address cap-based assistive devices [4], [5]. They were selected because together they cover location estimation and obstacle or object awareness,

the two capabilities that a wearable guide must combine.

Each study was compared on its task, platform, main approach, dataset, reported results and suitability for wearable assistive use. Only results stated by the original authors are quoted, and no study was re-implemented. The studies use different datasets and metrics, so their numbers describe each work on its own terms and are not directly comparable. The small number of studies is a limitation, and the review should be extended before firm conclusions are drawn.

III. VISUAL GEO-LOCALIZATION STUDIES

A. Global Image Geolocation with Satellite-Derived Attributes

Bianco et al. [1] estimate the location of a single ground-level image without GPS or other metadata. They introduce the Recall vs Area (RvA) metric, which reports the fraction of images whose true location falls inside the highest-scoring predicted regions as a function of the total area searched, so that discontinuous predictions can be assessed. The proposed method ensembles two existing geolocation models, GeoEstimation and GeoCLIP, with attribute predictors trained to recognise population density (LandScan) and land cover (ESA) from ground-level images. Their probability maps are combined by an element-wise product on a common latitude-longitude grid.

On the Street View set, balanced across population-density classes, adding the LandScan predictor raised GeoCLIP recall at the area budget equivalent to a 200 km radius from 44.70% to 54.42% [1]. Gains were largest in non-urban areas. However, recall on the Im2GPS3k set fell at small area budgets for GeoEstimation (39.42% to 35.08% at the 25 km-equivalent budget), top-1 distance accuracy changed little, and the common grid has a resolution of about 4 km, so street-level answers cannot be returned [1].

B. UAV Cross-View Geo-Localization

Wang and Zheng [2] localise a UAV in satellite-denied environments by matching its aerial image to satellite reference patches. A multi-scale reference library is built for different flight altitudes, with 50% overlap between patches, and a Vision Transformer (ViT) backbone is extended with a Multi-Scale and Long-range Attention (MSLA) module. On

University-1652 the method reached R@1 of 87.40% (drone to satellite) and 92.15% (satellite to drone), and flight tests with a DJI Mavic 3 at 80–220 m confirmed stable matching across altitudes [2].

The accuracy has a computational price: the model has 114.58 M parameters and takes 70 ms per image, compared with 89.00 M parameters and 55 ms for the ViT baseline, on the NVIDIA RTX 4090 GPU used in the experiments [2]. The authors name model efficiency for onboard deployment as future work [2].

C. Mamba-Based Deep Hashing

Liu et al. [3] match street-view images with a limited 90° field of view (FoV) to remote-sensing images. They replace the ViT with Vision Mamba (ViM), whose computation grows linearly rather than quadratically with sequence length, and add a hash encoder that converts features into binary codes compared by Hamming distance. On the CVACT dataset the method reached R@1 of 36.55%, above TransGeo (28.16%) and BEV-CV (34.15%), although its R@1% of 88.57% was below BEV-CV (94.51%) and S6G (90.15%) [3]. Inference time was 0.43 ms per batch against 22.75 ms for TransGeo on an RTX 2080 Ti GPU [3]. The absolute R@1 is still modest, and a remote-sensing reference database is required.

IV. CAP-BASED ASSISTIVE DEVICES

A. Smart Hat with Deep Learning

Almurayziq et al. [4] propose a smart hat for blind persons in smart cities, built around a virtual reality (VR) camera, a microphone and a speaker connected to a smartphone application. Objects are detected with You Only Look Once version 3 (YOLOv3) trained on the Common Objects in Context (COCO) dataset, a convolutional neural network (CNN) assigns them to three classes (traffic, objects and animals), and a voice-activated personal assistant (VAPA) with text-to-speech (TTS) speaks to the user. The architecture also uses cloud storage and external weather and traffic sources [4].

The evaluation covers four example images, with runtimes of 3.2–7.8 s per image and a qualitative ‘high’ confidence label. No precision, recall or other accuracy metrics are reported, and the authors describe the work as in progress [4]. No positioning or navigation module is described.

B. Ultrasonic Smart Cap

Rahman et al. [5] describe a low-cost cap built from an Arduino UNO, ultrasonic sensors, an NE555 timer circuit, a potentiometer for setting the detection range, a buzzer, a tactile switch and a 3.7 V lithium battery. The ultrasonic sensor has a stated range of 2–400 cm, and the buzzer tone changes with the distance of the obstacle [5]. The device is simple, inexpensive and easy to wear.

It warns of the presence of an obstacle but does not identify it, and the paper reports no quantitative detection accuracy, latency, battery life or user-study results [5]. The authors themselves identify the lack of seamless, real-time, user-friendly interaction as the remaining gap [5].

V. EVALUATION AND EFFICIENCY CONSIDERATIONS

The reviewed studies measure different things. Reference [2] uses Recall@K and average precision, [3] uses recall at K, [1] uses great circle distance and the RvA curve, [4] reports runtime per image, and [5] reports no measurements. None reports results on embedded hardware, and none measures battery life, false alarm rate or navigation success, which matter to a person walking with the device.

Two ideas are nevertheless useful for wearable design. First, RvA [1] treats a location estimate as a region whose size can be traded against recall, which is a natural way to express localization uncertainty to a downstream guidance module. Second, the binary-code retrieval of [3] shows that matching cost can be reduced greatly, which is relevant for devices with limited computing power. Conversely, the 70 ms per image and 114.58 M parameters of [2], and the 3.2–7.8 s per image of [4], show that accuracy and latency are in tension and must be reported together.

VI. LITERATURE REVIEW

The five studies fall into four research directions.

Global and regional geolocation: Study [1] improves image geolocation for under-sampled non-urban areas by adding satellite-derived attributes to existing models, and proposes an area-based metric.

Cross-view retrieval: Studies [2] and [3] match ground or aerial images against satellite imagery, the first

emphasising accuracy through multi-scale attention and the second emphasising speed through Mamba and hashing.

Vision-based assistive wearables: Study [4] combines object detection, classification and voice interaction in a hat, using a smartphone and cloud services.

Sensor-based assistive wearables: Study [5] uses ultrasonic ranging and a buzzer in a low-cost cap.

Overall, location estimation, object recognition and proximity sensing have each been studied, but not combined in one wearable that is evaluated under walking conditions.

VII. COMPARATIVE DISCUSSION AND RESEARCH GAPS

Object detection with voice output in a hat [4] and ultrasonic proximity warning in a cap [5] have already been demonstrated, and multi-scale and efficient retrieval have been studied for geo-localization [2], [3].

These capabilities should therefore not be presented as unexplored. The opportunity lies in integrating them on one wearable and evaluating them under the constraints of walking users.

The comparison is presented in Table I. Based on the reviewed studies; the main gaps are:

1) *No evaluation for wearable, pedestrian use:*

The geo-localization studies [1]– [3] use benchmark images from vehicles, aircraft or archived panoramas, not head-worn camera streams from walking users.

2) *Coarse or reference-dependent positioning:*

The method in [1] works on a grid of about 4 km, while [2] and [3] require stored satellite or remote-sensing reference data and give modest street-view accuracy (R@1 of 36.55% in [3]). None alone can guide a pedestrian.

3) *High computational cost and no embedded evidence:*

Results in [2] and [3] are obtained on desktop GPUs, and [4] reports multi-second runtimes. No study demonstrates real-time operation on a low-power single-board computer.

4) Limited perception in assistive caps:

The cap in [5] detects only proximity, and the hat in [4] covers three object classes. Neither reads text, recognises familiar people or describes the scene.

5) No sensor fusion or prioritised feedback:

Neither [4] nor [5] combines several sensors, and neither decides which of several simultaneous messages is most urgent.

6) Weak quantitative and user-centred evaluation:

The evaluation in [4] uses four example images and [5] reports none, and the geolocation studies do not measure user outcomes.

7) Limited safety and privacy provisions:

Neither cap includes fall detection or emergency alerts, and [4] depends on a smartphone and cloud storage.

TABLE I Comparison of the Reviewed Studies

Ref.	Task and platform	Main approach	Dataset and reported results	Limitation for wearable assistive use
[1]	Global single-image geolocation without metadata; software evaluation	GeoEstimation and GeoCLIP ensembled with population-density and land-cover predictors; RvA metric	Im2GPS3k and Street View (balanced). GeoCLIP + LandScan on Street View: recall 44.70% to 54.42% at 200 km-equivalent area	About 4 km grid; regional-scale gains; little top-1 improvement; no embedded evaluation
[2]	UAV localization in satellite-denied areas; RTX 4090 GPU	Multi-scale satellite patch library; ViT with MSLA module	University-1652: R@1 87.40% (drone to satellite), 92.15% (satellite to drone); flights at 80–220 m	114.58 M parameters, 70 ms per image; needs stored satellite map; aerial viewpoint
[3]	Street-view (90° FoV) to remote-sensing retrieval; RTX 2080 Ti GPU	Twin ViM backbone with hash encoder; Hamming-distance retrieval	CVACT: R@1 36.55%, R@1% 88.57%; 0.43 ms vs 22.75 ms per batch (TransGeo)	Modest R@1; GPU timing only; reference database required
[4]	Smart hat for blind users in smart cities; smartphone and cloud	VR camera; YOLOv3 on COCO; CNN classes; voice assistant and TTS	Four example images; 3.2–7.8 s per image; no accuracy metric	Multi-second latency; no positioning module; cloud dependence; work in progress
[5]	Obstacle warning cap; Arduino UNO	Ultrasonic ranging; NE555 timer; potentiometer range setting; buzzer	Hardware prototype; no quantitative results reported	Proximity alert only; no recognition, localization or evaluation data

VIII. PROPOSED GEOVISION CAP AND HOW IT ADDRESSES THE GAPS

A. System Overview

GeoVision Cap is a planned wearable Edge AI system for visually impaired users. Its hardware comprises a Raspberry Pi 5 (8 GB), a Raspberry Pi Camera Module 3, an optional Google Coral or Hailo AI accelerator, a VL53L1X Time-of-Flight (ToF) sensor, ultrasonic sensors, a GPS module, an MPU6050 inertial measurement unit (IMU), a ReSpeaker microphone array, bone-conduction headphones and a

rechargeable battery pack. The estimated prototype cost is INR 40,300.

The software stack uses Python with OpenCV, PyTorch, TensorFlow Lite and ONNX Runtime, PaddleOCR for text, Whisper for speech recognition, Coqui TTS for speech output, and a lightweight YOLO model for detection.

Camera and sensor data pass through Edge AI processing, sensor fusion and a decision engine before reaching the user as voice guidance (Fig. 1).

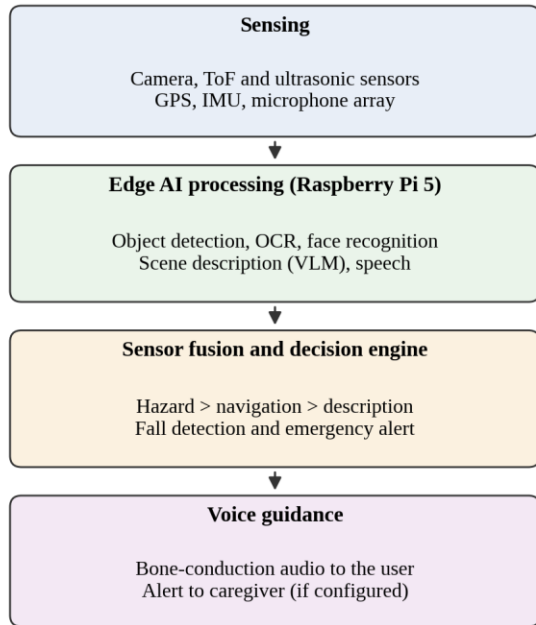


Fig. 1. Planned processing pipeline of GeoVision Cap.

B. Planned Functions

The planned functions are: object detection (persons, vehicles, doors, chairs, poles and similar); optical character recognition (OCR) for books, menus, currency, documents, medicine labels and street signs;

recognition of familiar faces; scene description by a vision-language model (VLM) in short natural sentences; spoken navigation using GPS and visual cues; voice commands such as asking what is around the user; and an emergency function that detects falls, sends GPS coordinates and alerts caregivers, subject to user configuration. Processing is intended to run mainly on the device, with cloud services optional for advanced reasoning.

C. Response to the Identified Gaps

Table II maps each gap to a planned design response and the measure that will be used to test it. The responses are design intentions; none has yet been validated.

On the geographic aspect of the name, GPS and the IMU are intended to be the primary position source, because the methods in [1] are too coarse and those in [2] and [3] too reference-dependent to replace them. The visual geo-localization ideas reviewed here are considered only as a coarse prior or fallback, for example matching against a small locally stored set of familiar places using compact binary codes as in [3], with the uncertainty expressed in the area-budget manner of [1]. Whether this improves guidance on a wearable is an open question to be tested.

TABLE II Research Gaps and the Planned GeoVision Cap Response

No.	Research gap	Planned GeoVision Cap response	Planned measure
1	No evaluation for wearable, pedestrian use	Collect walking sequences with the cap camera and test localization aids against a GPS reference log	Localization error against GPS log
2	Coarse or reference-dependent positioning	GPS and IMU as primary position source; visual geo-cues only as coarse prior or fallback with stated uncertainty; ToF and ultrasonic sensors for near-range hazards	Navigation success rate
3	High computational cost, no embedded evidence	Edge-first design on Raspberry Pi 5 with optional accelerator; lightweight YOLO model; quantised models; compact binary-code matching for a small local reference set	Response time; battery life
4	Limited perception in assistive caps	Detection, OCR, face recognition and VLM scene description combined on one device	Detection, OCR and face recognition accuracy
5	No sensor fusion or prioritised feedback	Decision engine ranks hazards above navigation above description; short bone-conduction messages that leave ambient sound audible	False alarm rate; user satisfaction
6	Weak quantitative and user-centred evaluation	Full metric set (see Section IX); comparison with an ultrasonic-only baseline like [5] and a smartphone-and-cloud pipeline like [4] on the same routes	All measures above
7	Limited safety and privacy provisions	IMU-based fall detection; GPS alert to caregivers if configured; local processing by default, cloud optional	Fall-detection recall; false alarms

IX. FUTURE RESEARCH DIRECTION

The next step is to build the prototype and measure latency and power consumption on the Raspberry Pi 5 with and without an accelerator. The system will then be evaluated on detection accuracy, OCR accuracy, face recognition accuracy, speech recognition accuracy, response time, battery life, user satisfaction, navigation success rate and false alarm rate. Localization uncertainty will additionally be reported as recall against search area, following [1].

Comparisons with an ultrasonic-only baseline [5] and a smartphone-and-cloud pipeline [4] should use identical indoor and outdoor routes. Trials with visually impaired participants require informed consent and appropriate ethical approval. Further work may include indoor simultaneous localization and mapping (SLAM), multilingual voice support and edge-based vision-language models. The review itself should also be extended to a larger set of studies.

X. CONCLUSION

The reviewed studies show that accurate geo-localization and low-cost obstacle warning have each been achieved, but in separate settings: the former on powerful hardware with archived imagery, the latter with minimal sensing and little evaluation. Seven gaps follow from this separation, concerning pedestrian evaluation, positioning precision, embedded efficiency, perception range, sensor fusion, evaluation quality, and safety and privacy.

GeoVision Cap is proposed as an integrated, edge-first wearable aimed at these gaps. Its value will depend on measured performance with real users, which this paper does not yet provide. The planned prototype and evaluation are the logical continuation of this review.

REFERENCES

- [1] M. J. Bianco, D. Eigen, and M. Gormish, "Enhancing worldwide image geolocation by ensembling satellite-based ground-level attribute predictors," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, 2025, pp. 535–543, doi: 10.1109/WACVW65960.2025.00065.
- [2] X. Wang and L. Zheng, "UAV cross-view geo-localization based on multi-scale partitioning and attention-enhanced transformer," in *Proc. 2025 Int. Conf. Signal Process., Comput. Netw. Commun. (SPCNC)*, 2025, pp. 512–516, doi: 10.1109/SPCNC68200.2025.11406231.
- [3] J. Liu, K. He, and P. Ren, "Mamba-based deep hashing for geo-localization of street view images," in *Proc. 2025 Int. Conf. New Trends Comput. Intell. (NTCI)*, 2025, pp. 458–462, doi: 10.1109/NTCI67886.2025.11308194.
- [4] T. S. Almurayziq, N. Alotibi, G. Alshammari, A. Alshammari, and M. Alsaffar, "Smart and guide hat for blind persons in smart cities using deep learning," *J. Adv. Inf. Technol.*, vol. 14, no. 6, pp. 1214–1220, 2023, doi: 10.12720/jait.14.6.1214-1220.
- [5] M. Rahman, M. R. Islam, and S. M. A. Islam, "Smartly controlled cap for blind individuals," *J. Sens. Res. Technol.*, vol. 7, no. 1, pp. 29–38, 2025, doi: 10.5281/zenodo.14942588.