OpsisVision: A Multimodal AI Assistance System for the Blind and Visually Impaired with Hybrid Edge-Cloud Architecture

Author: Grigorios Tsinaforniotis, Protos AI Agency (Single Member Private Company)

Date: December 29, 2025

Abstract

Assistive technologies for the blind and visually impaired have reached enormous potential through advances in artificial intelligence (AI) and wearable computing systems. However, many existing solutions remain limited by high costs, proprietary hardware, or dependence on pure cloud services. This paper presents **OpsisVision**, a multimodal AI assistance system based on a low-cost, 3D-printed glasses frame with standard hardware. The system combines local real-time obstacle detection using YOLOv11 with precise depth measurement through stereoscopic vision (Stereo Vision). For deeper semantic understanding of the environment, a hybrid architecture is used that combines local object detection with visual analysis by a Large Language Model (GPT-4o) in the cloud. Interaction is entirely via voice commands in Greek, enabled by local wake-word detection (Vosk) and cloud-based speech recognition (Whisper). We describe the system's architecture, its implementation on an NVIDIA Jetson Orin, and the results of a successful test run. A detailed cost analysis shows that OpsisVision, with estimated hardware costs of approximately 470 euros (8GB cloud version) to 2,200 euros (64GB offline version), is significantly cheaper than commercial alternatives such as Envision Glasses ($3,500) or OrCam MyEye Pro ($6,000). Furthermore, we discuss the ethical implications, particularly with regard to the EU General Data Protection Regulation (GDPR) and the European Accessibility Act (EAA) 2025. OpsisVision demonstrates a scalable and cost-effective approach that can be evolved from a pure cloud prototype to a fully offline-capable system, thereby significantly increasing the accessibility and independence of the blind and visually impaired.

---

1. Introduction

1.1 Background and Motivation

According to estimates by the World Health Organization (WHO), at least 2.2 billion people worldwide live with a vision impairment, of whom at least 1 billion have a vision impairment that could have been prevented or has not yet been treated [7]. In the European Union, approximately 30 million people are blind or visually impaired [8]. For many of these individuals, especially those with severe vision impairment or blindness, navigating in unfamiliar or dynamic environments poses a daily challenge.

Traditional aids such as the white cane or guide dogs, while indispensable, provide only limited information about the environment. The white cane primarily detects obstacles at ground level and in the immediate vicinity, while guide dogs, although able to navigate more intelligently, require intensive training and are not accessible to everyone. Both aids provide no information about the semantic meaning of a scene—such as whether a café, a pharmacy, or a bus stop is nearby.

The rapid advances in artificial intelligence, particularly in the fields of computer vision and natural language processing, have opened up new possibilities for assistive technologies. Modern deep learning models can detect objects in real time, describe scenes, and even answer complex questions about visual content. At the same time, powerful edge computing platforms like the NVIDIA Jetson series enable the execution of such models on portable devices.

1.2 Problem Statement

Despite these technological advances, existing solutions for the blind and visually impaired often remain inaccessible. The main reasons for this are:

High Costs: Commercial products such as Envision Glasses (approx. $3,500) or OrCam MyEye Pro (approx. $6,000) are unaffordable for many affected individuals, especially in countries with lower incomes [9].

Proprietary Hardware: Many systems are tied to specific, non-interchangeable hardware, which complicates repairs and increases dependence on the manufacturer.

Cloud Dependency: Most modern AI glasses require a constant internet connection, which limits their use in areas without a stable connection and raises significant privacy concerns.

Language Limitations: Many systems are primarily focused on English, which limits usability for speakers of other languages.

1.3 Contribution of this Work

OpsisVision was developed to close this gap. Our goal was to create a system that meets the following criteria:

Cost-effective and Reproducible: Use of standard hardware (Jetson, webcams) and a 3D-printed frame, with estimated total costs of approximately 470-2,200 euros—significantly cheaper than commercial alternatives.

Multimodal: Combination of object detection, depth measurement, and semantic analysis for a comprehensive understanding of the environment.

Hybrid: Intelligent use of local (edge) and cloud resources for an optimal balance of speed and intelligence.

Natural Interaction: Full voice control in the user's native language (here: Greek).

Scalable: A clear development path from a cloud-assisted prototype to a fully autonomous offline version.

Privacy-Compliant: Architecture that allows for minimization of data sent to the cloud and can prospectively work completely offline.

This paper describes the current prototype of OpsisVision, which will be presented to the public in Greece in January 2026. We explain the system architecture, the hardware and software components, present the results of a successful test run, and discuss the ethical and regulatory implications.

---

2. Related Work

The development of assistive technologies for the blind is an active field of research with a long history. In this section, we provide a comprehensive overview of existing commercial products and academic research.

2.1 Commercial Systems

The following table provides an overview of the main commercial AI glasses for the blind and visually impaired:

Product	Manufacturer	Price	Main Features	Cloud Dependency
Envision Glasses	Envision	$3,500 + $200/year	Text recognition, face recognition, scene description	Yes
OrCam MyEye Pro	OrCam	$6,000	Text recognition, face recognition, product recognition	Partial
OrCam Read Smart	OrCam	$1,990	Text recognition	No

Product	Manufacturer	Price	Main Features	Cloud Dependency
Ally Solos	Ally	$399 + subscription	Scene description, navigation	Yes
Be My Eyes + Meta	Meta/Be My Eyes	Glasses price + app	LLM-based scene description	Yes

Envision Glasses [1] are based on the Google Glass Enterprise Edition 2 and offer a wide range of functions, including Instant Text (immediate text recognition), Scan Text (for longer documents), Call an Ally (connection to a sighted helper), and Describe Scene (AI-based scene description). The glasses are lightweight and unobtrusive but require a constant internet connection for most functions.

OrCam MyEye [10] is a small device that can be attached to any pair of glasses. It offers text recognition, face recognition, and product recognition. Unlike Envision, OrCam works partially offline, which improves privacy but limits functionality.

The integration of **Be My Eyes into the Ray-Ban Meta Glasses** [2] represents a paradigm shift. By using Large Language Models (LLMs), these glasses can not only recognize objects but also answer complex questions about the environment. However, the dependence on Meta's cloud infrastructure raises questions about data privacy and long-term availability.

2.2 Academic Research

Academic research has produced numerous innovative approaches that often serve as the basis for commercial products.

Early Work (2010-2016): The pioneering work of Poggi and Mattoccia (2016) demonstrated the potential of 3D vision and deep learning for the navigation of the blind [4]. Their system used a stereo camera and a Convolutional Neural Network (CNN) for obstacle detection and was tested with blind users.

Binocular Systems (2017-2020): Jiang et al. (2018) developed a wearable system with binocular sensors that enables precise depth perception [3]. Their work emphasized the importance of real-time processing for user safety.

Deep Learning Approaches (2019-2022): Lin et al. (2019) presented a system based on semantic segmentation [6]. By classifying each pixel in the image, the system can not only detect obstacles but also determine their type (e.g., stairs, curb, vehicle).

LLM Integration (2023-2025): The latest development is the integration of Large Language Models into assistance systems. Waisberg et al. (2024) discussed the potential of Meta Smart Glasses with LLM integration for the visually impaired [11]. These systems can answer natural language questions and provide context-related information.

Current Developments (2025): Gao et al. (2025) presented a wearable obstacle avoidance device weighing about 400 grams [5]. Ahmad (2025) developed a virtual assistance system that combines object detection with position and distance estimation [12].

2.3 Research Gap and Positioning of OpsisVision

Despite the multitude of research papers and commercial products, a clear gap remains: there is a lack of a system that is simultaneously cost-effective, multimodal, hybrid (edge + cloud), linguistically flexible, and privacy-friendly.

OpsisVision addresses this gap through its unique architecture:

Feature	Envision	OrCam	Be My Eyes/Meta	OpsisVision
Cost	High	Very High	Medium	Low
Depth Measurement	No	No	No	Yes (Stereo)
LLM Integration	Yes	No	Yes	Yes
Offline Capable	No	Partial	No	Planned
Open Source	No	No	No	Yes
Non-English Languages	Limited	Limited	Limited	Full

---

3. System Architecture

OpsisVision consists of a hardware unit (the glasses) and a software pipeline that runs on an NVIDIA Jetson Orin Nano. The architecture is modular to allow for easy maintenance and future extensions.

3.1 Hardware Components

The hardware components were selected with cost-effectiveness, availability, and performance in mind.

Component	Specification	Purpose	Estimated Cost
Compute Unit	NVIDIA Jetson Orin (8GB/16GB/64GB)	Local AI processing	€250-€2,000
Cameras	2x 4K USB webcams	Stereoscopic image capture	~€80
Frame	3D-printed PLA	Mount for cameras	~€10
Microphone	USB microphone	Voice input	~€20
Speaker	Bone conduction headphones	Voice output	~€50
Power Supply	Anker Powerbank (USB-C PD, 45W)	Mobile power supply	~€60
Total			**~€470-€2,220**

Figure 1: The OpsisVision prototype with 3D-printed frame and stereo cameras.

NVIDIA Jetson Orin: This edge computing platform offers impressive performance of up to 275 TOPS (Tera Operations Per Second) in its most powerful variant. For OpsisVision, three configurations are available: the 8GB version works excellently in cloud-hybrid mode, the 16GB version enables offline operation, and the 64GB version offers optimal offline performance with significantly higher speed. All variants support CUDA, cuDNN, and TensorRT, which enables the execution of optimized deep learning models.

Stereo Cameras: The use of two 4K cameras with a baseline of about 6 cm allows for precise depth measurement through stereo vision. The high resolution ensures accurate object detection even at greater distances.

3D-Printed Frame: The frame was printed in PLA (Polylactic Acid), a cost-effective and environmentally friendly material. The design is modular, so cameras and electronics can be easily replaced.

3.2 Software Architecture

The software pipeline is a Python script that triggers a chain of events, starting with the detection of the wake word.

Step 1: Wake-Word Detection (Local)

A Vosk recognizer [13] runs continuously and listens for the Greek wake word "γυαλιά" (glasses). Vosk is an open-source speech recognition system that works completely offline. This ensures data privacy and minimizes power consumption, as no data is sent to the cloud until the user explicitly activates the system.

Step 2: Question Recording and Transcription (Cloud)

After activation, a 5-second audio recording is made and sent to the OpenAI Whisper API for transcription. Whisper [14] is a state-of-the-art speech recognition model that supports over 90 languages, including Greek with high accuracy.

Step 3: Image Capture and Obstacle Detection (Local)

Simultaneously, images are captured from the two 4K cameras. The left image is immediately processed by a local YOLOv11 model [15] to detect objects and their position (left, center, right). YOLOv11 (You Only Look Once, Version 11) is the latest version of the popular real-time object detection model and offers an excellent balance between speed and accuracy.

Step 4: Depth Measurement (Local)

A disparity map is calculated from the left and right images using the OpenCV StereoBM algorithm [16]. Disparity is the horizontal difference in the position of an object between the left and right images. The distance can be calculated from the disparity:

```
Distance = (Baseline × Focal Length) / Disparity
```

For each detected object, the median disparity in the bounding box is used to estimate the distance in centimeters.

Step 5: Scene Analysis (Cloud)

The left image is sent to the GPT-4o Vision API [17] to obtain a rich, semantic description of the entire scene. GPT-4o can not only name objects but also describe their relationships to each other, capture the atmosphere of a scene, and provide context-related information.

Step 6: Response Generation and Speech Output

The detected obstacles with their distances and the scene description from GPT-4o are combined into a coherent response. This is converted into Greek speech using the OpenAI TTS API with the voice "nova" and output via the speaker.

Figure 2: The author wearing the OpsisVision prototype.

3.3 Hybrid Architecture: Rationale and Advantages

The decision for a hybrid edge-cloud architecture is based on a careful consideration of latency, accuracy, cost, and data privacy.

Task	Processing	Rationale
Wake-Word	Local (Vosk)	Privacy, low latency, power saving
Object Detection	Local (YOLO)	Safety-critical, low latency required
Depth Measurement	Local (OpenCV)	Computationally intensive but efficient on GPU
Speech Recognition	Cloud (Whisper)	High accuracy for Greek required

Task	Processing	Rationale
Scene Analysis	Cloud (GPT-4o)	LLM too large for edge device
Speech Synthesis	Cloud (TTS)	Natural voice required

The safety-critical tasks (obstacle detection, depth measurement) are processed locally to ensure minimal latency. The cloud is only used for tasks that either require very high computational power (LLM) or benefit from the quality of cloud models (speech recognition, speech synthesis).

3.4 Scalability: From Prototype to Offline Version

OpsisVision offers flexible hardware options for different requirements and budgets:

Three configurations are available:

Version	RAM	Operating Mode	Response Time	Cost
Jetson Orin Nano 8GB	8GB	Cloud-Hybrid (Online)	~2-3 seconds	~€250
Jetson Orin 16GB	16GB	Offline	~5-8 seconds	~€1,000
Jetson Orin 64GB	64GB	Offline optimal	~2-3 seconds	~€2,000

The 8GB version works excellently in cloud-hybrid mode, where computationally intensive tasks (Whisper, GPT-4o) are processed in the cloud. The 16GB version enables full offline operation with a response time of approximately 5-8 seconds. For optimal offline performance with fast response times, the use of a Jetson Orin with 64GB RAM is recommended. This would allow more powerful local models to be run directly on the device:

Model	Cloud Version	Local Alternative	Requirements
Whisper	Whisper API	Whisper Large v3	~10GB VRAM
GPT-4o	GPT-4o API	Llama 3.2 Vision	~40GB RAM

Model	Cloud Version	Local Alternative	Requirements
TTS	OpenAI TTS	Piper TTS	~1GB RAM

With a 64GB Jetson, OpsisVision could work completely offline, which is of crucial importance for the independence and privacy of users.

---

4. Implementation and Results

4.1 Test Environment

To validate the functionality of the system, a test run was conducted in an indoor environment. The test environment was a living room with various pieces of furniture, electronic devices, and decorative items.

4.2 Test Results

The system was activated by the wake word and confronted with the question "What do you see?". The following log excerpt shows the system's response:

```
[15:07:08] [LOCAL] Listening for wake word...
```

```
[15:07:24] [LOCAL] Wake word "γυαλιά" detected!
[15:07:24] [SYSTEM] Voice output: "Ναι, πες μου;" (Yes, tell me?)
[15:07:25] [LOCAL] Listening for question (5s)...
[15:07:30] [CLOUD] Question transcribed: "τι βλέπεις" (What do you see?)
[15:07:30] [LOCAL] Capturing images...
[15:07:31] [LOCAL] Υπολογισμός βάθους... (Depth calculation...)
[15:07:32] [LOCAL] Εντοπισμός εμποδίων... (Obstacle detection...)
[15:07:33] [LOCAL] Obstacle detected: TV, position: front, distance: 7
cm
[15:07:34] [CLOUD] Ανάλυση σκηνής... (Scene analysis...)
[15:07:38] [CLOUD] Scene analysis received.
```

```
[15:07:39] [SYSTEM] Voice output: "Εμπόδια: tv μπροστά, περίπου 7
εκατοστά..."
```

Translation of the voice output:

> *"Obstacles: TV in front, approximately 7 centimeters. In the image, a room with a wooden desk in the foreground is visible. On the desk are objects such as a charger and*

> *stationery. In the background, there is a black sofa with cushions and a decorative painting on the wall. There is also a floor lamp and a potted plant. At the top of the image, part of a computer monitor is visible."*

4.3 Analysis of Results

The test run successfully demonstrates the entire processing chain:

Phase	**Latency**	**Result**
Wake-Word Detection	<100ms	Successful
Question Recording	5s (configured)	Successful
Transcription (Cloud)	~500ms	Correct
Image Capture	<100ms	Successful
Depth Calculation	~500ms	Successful
Obstacle Detection	~300ms	TV detected, 7cm
Scene Analysis (Cloud)	~4s	Detailed description
Speech Synthesis	~1s	Natural output
Total Latency	**~12s**	Acceptable for non-critical queries

The ability to provide both quantitative data (distance to obstacle: 7 cm) and qualitative descriptions (scene analysis) is one of the core strengths of OpsisVision.

---

5. Cost Analysis

A key goal of OpsisVision is to provide an affordable alternative to commercial products. In this section, we compare the costs of OpsisVision with existing solutions.

5.1 Hardware Costs

Component	OpsisVision	Envision	OrCam MyEye Pro
Compute Unit	€250-€2,000 (Jetson 8-64GB)	Included	Included
Cameras	~€80 (2x 4K)	Included	Included
Frame	~€10 (3D print)	Included	Clip-on
Audio	~€70	Included	Included
Power Supply	~€60	Included	Included
Total Hardware	**~€470-€2,220**	**~€3,200**	**~€5,500**

5.2 Running Costs

Cost Factor	OpsisVision	Envision	OrCam MyEye Pro
Cloud API (estimated)	~€5-10/month	Included	€0
Software Updates	€0 (Open Source)	$200/year	€0
Annual Costs	**~€60-€120**	**~€180**	**€0**

5.3 Total Cost of Ownership (3 Years)

System	Acquisition	3 Years Operation	Total
OpsisVision (8GB, Cloud-Hybrid)	€470	€180-€360	**€650-€830**
OpsisVision (16GB, Offline)	€1,220	€0	**€1,220**
OpsisVision (64GB, Offline fast)	€2,220	€0	**€2,220**

System	Acquisition	3 Years Operation	Total
Envision Glasses	€3,200	€540	**€3,740**
OrCam MyEye Pro	€5,500	€0	**€5,500**

OpsisVision is thus **2-8 times cheaper** than commercial alternatives, with comparable or even superior functionality (stereo vision for depth measurement).

---

6. Ethical Considerations and Regulatory Framework

The development of AI-powered assistive technologies raises important ethical and regulatory questions. In this section, we discuss the relevant aspects in the context of OpsisVision.

6.1 Data Privacy and GDPR

The EU General Data Protection Regulation (GDPR) imposes strict requirements on the processing of personal data [18]. For OpsisVision, the following aspects are relevant:

Data Minimization: OpsisVision follows the principle of data minimization. The wake-word detection is done completely locally, so no audio data is sent to the cloud until the user explicitly activates the system.

Purpose Limitation: The data sent to the cloud (audio, images) are used exclusively for immediate processing and are not stored.

Consent: The user consciously activates the system by the wake word, thereby implicitly giving their consent to data processing.

Right to Erasure: Since no data is permanently stored, the right to erasure is automatically fulfilled.

The planned offline version of OpsisVision would completely eliminate all data privacy concerns, as no data would leave the device.

6.2 European Accessibility Act (EAA) 2025

The European Accessibility Act (EAA), which came into force on June 28, 2025, sets common accessibility requirements for products and services in the EU [19]. The EAA requires, among other things:

Compatibility with Assistive Technologies: Products must be compatible with assistive technologies. OpsisVision is itself an assistive technology and thus fulfills the spirit of the EAA.

Accessibility: The EAA aims to break down barriers for people with disabilities. OpsisVision contributes directly to this goal by enabling a higher degree of independence for the blind and visually impaired.

Affordability: Although the EAA does not prescribe price caps, affordability is an implicit goal. OpsisVision's low price makes the technology accessible to a wider range of people.

6.3 Ethical Considerations

Autonomy and Dignity: OpsisVision was developed with the aim of promoting the autonomy and dignity of the blind and visually impaired. The system is intended to support the user, not to patronize them.

Bias in AI Models: The AI models used (YOLO, GPT-4o) were trained on large, but possibly not fully representative datasets. It is important to continuously check the system for bias, especially with regard to the recognition of objects and people of different cultures and backgrounds.

Dependence on Technology: There is a risk that users will become too dependent on the technology. OpsisVision should be seen as a supplement to, not a replacement for, traditional aids such as the white cane.

---

7. Discussion and Outlook

7.1 Strengths of OpsisVision

OpsisVision offers several unique advantages over existing solutions:

Stereo Vision: As one of the few systems, OpsisVision offers real depth measurement through stereo vision. This enables precise distance information, which is essential for navigation.

Hybrid Architecture: The intelligent combination of edge and cloud processing offers an optimal balance between latency, accuracy, and cost.

Language Flexibility: The support for Greek demonstrates the system's adaptability to different languages and cultures.

Open Source: The open approach allows other developers and researchers to improve the system and adapt it to their needs.

Cost-Effectiveness: With estimated costs of under €500, OpsisVision is significantly cheaper than commercial alternatives.

7.2 Limitations

Hardware Design: The current prototype is still bulky and wired. Future versions will focus on miniaturization and the integration of components into a more aesthetic glasses frame. The goal is a slim, unobtrusive design, as shown in Figure 3.

Figure 3: Vision for the final product design of OpsisVision.

Cloud Dependency: The current prototype relies on the cloud for some functions. This limits its use in areas without a stable internet connection.

Latency: The total latency of about 12 seconds is acceptable for non-critical queries, but too high for real-time navigation. Future optimizations could reduce the latency.

User Tests: Extensive tests with blind and visually impaired users are still pending.

7.3 Future Work

Miniaturization: Development of a more compact, aesthetic design.

Offline Capability: Implementation of a fully offline-capable version with a 64GB Jetson.

User Tests: Conducting extensive tests with blind and visually impaired users.

Extended Functions: Integration of text recognition (OCR), face recognition, color recognition, and indoor navigation.

Investor Search: To finance further development and commercialization.

---

8. Conclusion

OpsisVision is more than just a concept; it is a functioning prototype that demonstrates the feasibility of a cost-effective, multimodal, and voice-controlled AI glasses for the blind. By combining local real-time processing and cloud intelligence, the system offers a robust and practical solution to the daily challenges of people with visual impairments.

With estimated costs of approximately €470-€2,220, OpsisVision is 2-8 times cheaper than commercial alternatives, with comparable or even superior functionality. The hybrid architecture enables an intelligent balance between latency, accuracy, and data privacy, while the open approach encourages further development by the community.

The public presentation of the prototype in January 2026 in Greece will be an important milestone. We believe that OpsisVision can pave the way for a new generation of affordable and adaptable assistive technologies that can improve the lives of millions of people with visual impairments worldwide.

References

[1] Envision. "Envision Glasses - AI for your eye." https://www.letsenvision.com/glasses

[2] Be My Eyes. "Be My Eyes on Ray-Ban Meta Glasses." https://www.bemyeyes.com/be-my-eyes-smartglasses

[3] Jiang, B., et al. (2018). "Wearable vision assistance system based on binocular sensors for visually impaired users." *IEEE Internet of Things Journal*.

[4] Poggi, M., & Mattoccia, S. (2016). "A wearable mobility aid for the visually impaired based on embedded 3D vision and deep learning." *IEEE Symposium on Computers and Communications*.

[5] Gao, Y., et al. (2025). "A wearable obstacle avoidance device for visually impaired." *Nature Communications*.

[6] Lin, Y., et al. (2019). "Deep Learning Based Wearable Assistive System for Visually Impaired People." *ICCV Workshop on Computer Vision Workshops*.

[7] World Health Organization. (2021). "World report on vision." https://www.who.int/publications/i/item/9789241516570

[8] European Blind Union. "Facts and figures about blindness and visual impairment." https://www.euroblind.org/

[9] Florida Reading. (2025). "OrCam vs. Envision vs. Ally Solos: An In-Depth Guide." https://floridareading.com/

[10] OrCam. "OrCam MyEye." https://www.orcam.com/

[11] Waisberg, E., et al. (2024). "Meta smart glasses—large language models and the future for assistive glasses for individuals with vision impairments." *Eye (Nature)*.

[12] Ahmad, F. (2025). "Virtual Assistance System for Visually Impaired people." *Procedia Computer Science*.

[13] Alpha Cephei. "Vosk Speech Recognition Toolkit." https://alphacephei.com/vosk/

[14] OpenAI. "Whisper." https://openai.com/research/whisper

[15] Ultralytics. "YOLOv11." https://docs.ultralytics.com/

[16] OpenCV. "Stereo Correspondence using Block Matching." https://docs.opencv.org/

[17] OpenAI. "GPT-4o." https://openai.com/gpt-4o

[18] European Commission. "General Data Protection Regulation (GDPR)." https://gdpr.eu/

[19] European Commission. "European Accessibility Act." https://ec.europa.eu/social/main.jsp?catId=1202