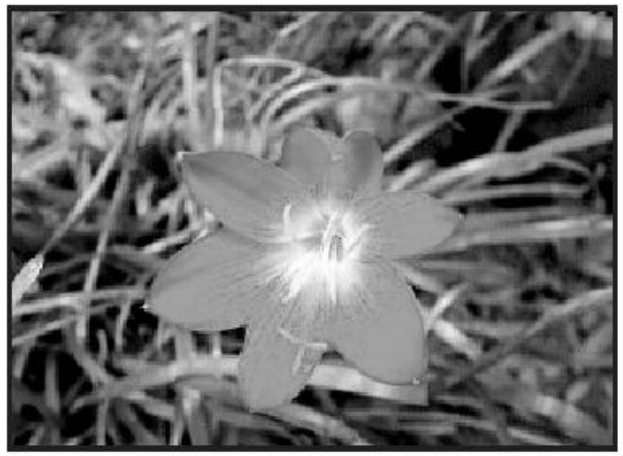
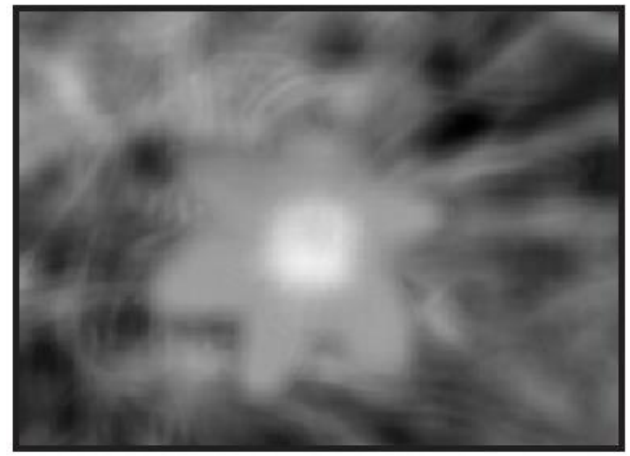
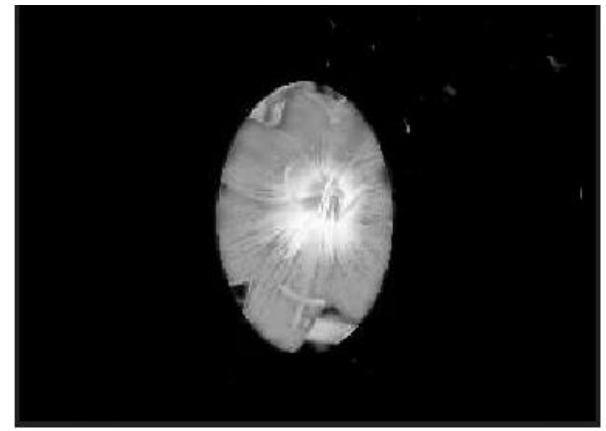
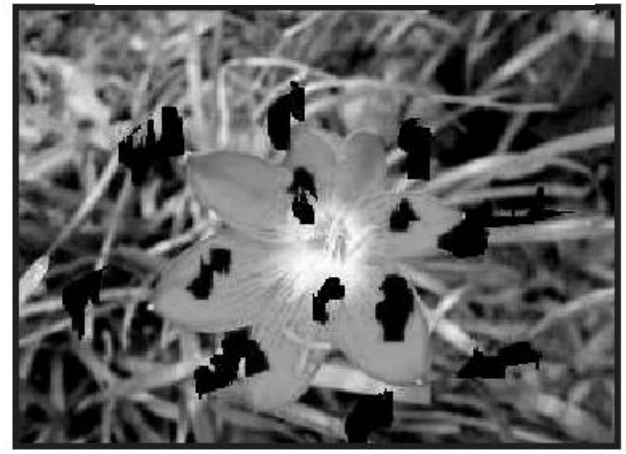








Object Search and Retrieval

- People with visual impairments face difficulties in **locating** and **retrieving** objects
- Limitation in independence by requiring a third party to help
- Solution: Blind Assistant Systems

NaviSense: Motivation

- Address weaknesses of existing systems:
  - Dependency on Human Assistance (BeMyEyes)
  - Dependency on Wearables (Ray-Ban MetaGlasses)
  - Limited Object Categories (ThirdEye)
  - Lack of precise guidance (WorldScribe)
- **Goal:** Provide Open-World Object Detection pipeline that identifies and guides users to objects

NaviSense: Components

- **Conversational Voice Interface:**
  - Apple Speech-to-Text
  - OpenAI GPT-4o-mini: interpreting user intent
- **3D Object Detection:**
  - Moondream 2B: an open-world VLM for 2D Object Detection
  - ARKit and LiDAR depth data to convert 2D location into spatial point
    - ➢ Continuously tracked and updated locally
- **Multimodal Guidance Feedback (Audio-Haptic):**
  - Audio Feedback: Directional voice prompts
  - Haptic Feedback: Distance-based vibrations

NaviSense: Evaluation

- User Study: 12 PVI participants
- Three target objects (3 trials per object), 4 tier shelf setup, 5 feet distance

Metric	Be My AI	Meta Glasses	NaviSense (ours)
Search Time (s)	48.23 ± 19.64	21.45 ± 14.58	**15.86 ± 5.65**
Guidance Time (s)	15.97 ± 12.85	19.35 ± 14.23	**15.89 ± 6.04**
Total Time (s)	64.19 ± 25.01	40.80 ± 20.83	**31.75 ± 8.11**
Undesired Objects	2.65 ± 2.07	4.37 ± 2.69	**0.52 ± 0.85**
Accuracy (%)	85.71	55.88	**95.37**

(1): Fig. 4. User Study Experiment Setup.

NaviSense: Strengths

- Precise guidance
- Intuitive interface
- Potential for enhanced privacy compared to e.g. BeMyEyes
- 24/7 availability

NaviSense: Weaknesses

- Initiating interactions verbally
  - ➢Not suitable for every environment
- Usage of two LLMs: one for interpretation, other for Object Detection
- iOS only
- Long search time (15 seconds average)
- Missing feature: Personal Object Retrieval
- Integration into existing assistive technologies unclear (VoiceOver, gesture control, …)

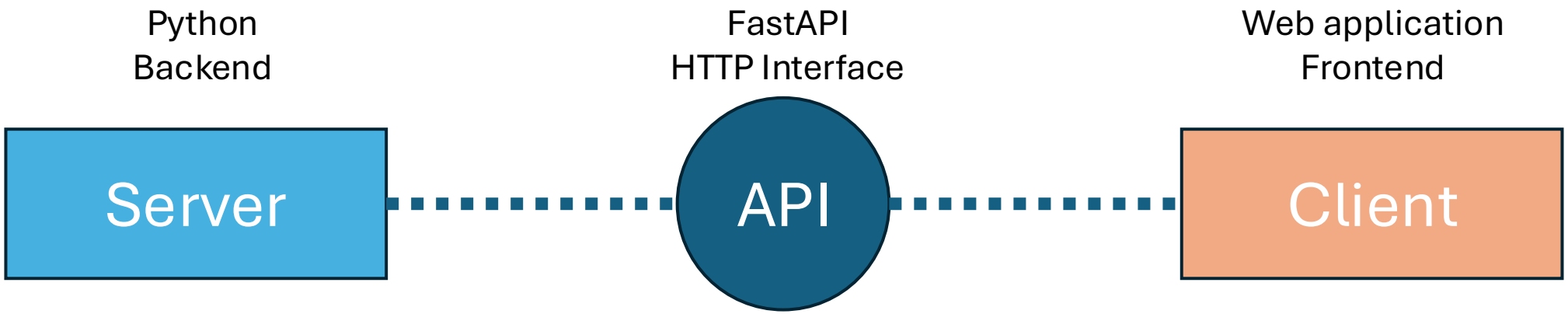
SeeEverything

By Elias Vollrath and
Annika Unmüßig

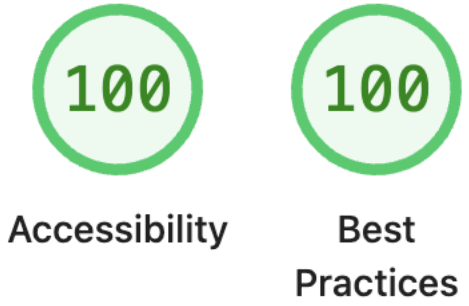
Goals

- Providing a tool that:
  - Operates in real-time
  - Combines Generic Object Detection and Personal Object Finding
  - Provides Guidance via Audio Feedback
  - Runs on every platform
  - Integrates well with mature accessibility frameworks
  - Uses little resources while running

Tool Architecture



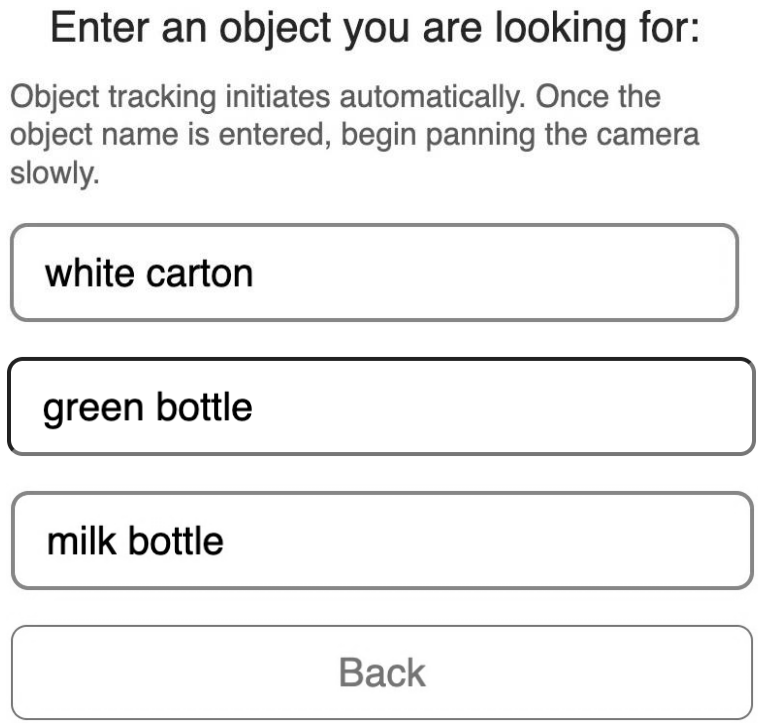
Frontend Implementation



- HTML, CSS, JavaScript
- Responsive Design
- Accessible Design
  - Lighthouse Accessibility score avg. 99%
  - Audio feedback during UI navigation
  - Usage of ARIA roles
  - High-contrast interface
- Hardware accelerated using modern APIs (WebGPU, WASM)

Generic Object Detector: OD Architecture

- YOLO World: Open World Object Detector
- Enables real-time detection of any object based on descriptive text



Depth Estimation + Guidance

- Depth estimation:
  - “Depth Anything”: Supplies relative per-pixel depth information
  - Evolution of bounding box size

- Guidance via “parking sensor”:
  - Estimated depth decreased over time?
  - Bounding box size increased over time?

  => Object got closer: increase beep/vibration frequency

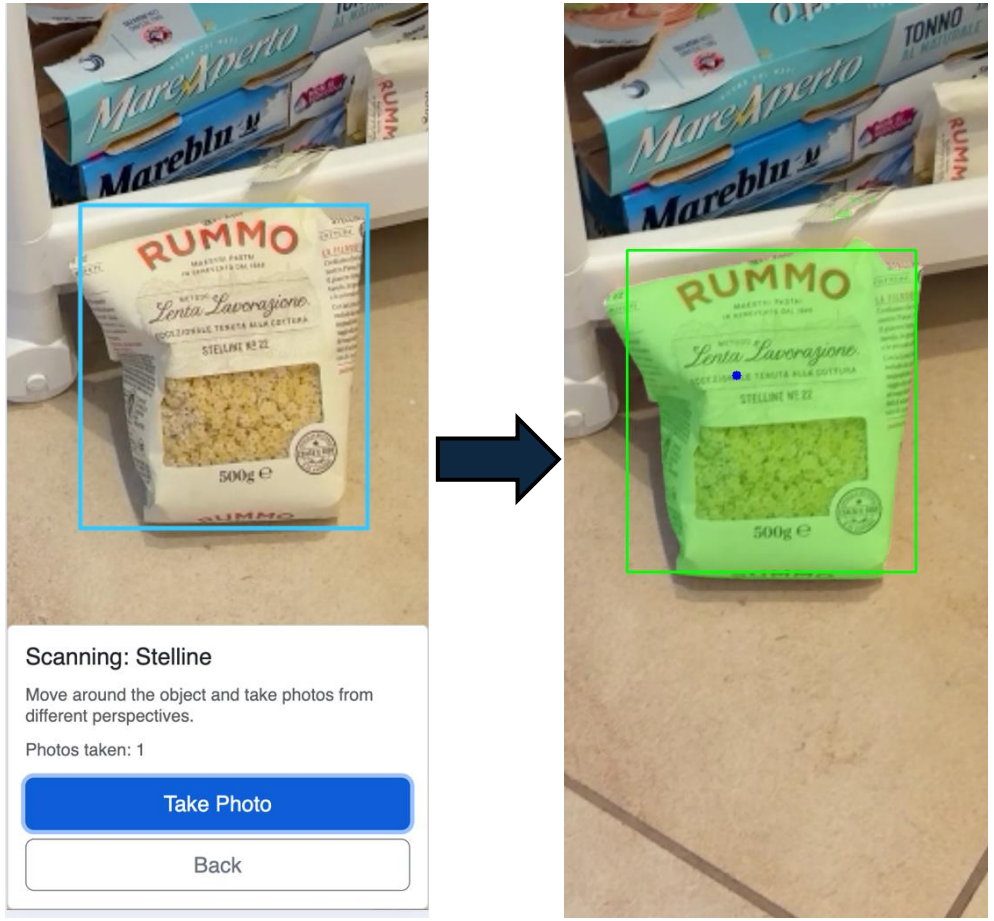
Personal Object Finder

- Goal: develop system that makes it possible for users to save their personal items and help them retrieve them later
- Related Work:
  - Recog: asks users to provide 10+ images of personal objects and trains CNN on those images
    - ➢ No guidance provided to objects
    - ➢ Requires blank background when taking images
    - ➢ Training inefficient (3 hours), costly
- How to improve this?

Personal Object Scanner: Models

- SAMv2 (Segment Anything Model 2) for object segmentation
- Object tracking based on the Lucas-Kanade algorithm
- DINOv2 for feature extraction of the object
- FAISS database

Personal Object Scanner: Logical Flow



1. Input new personal Object name
2. Align object in the center of the screen (marked with dot)
3. Start tracking
4. Take photo of the object
5. SAMv2 segmentation, and DINOv2 feature embedding
6. Storing in FAISS database

Personal Object Finder

- FastSAM for automatic mask proposal
- Features of each mask get calculated and compared with feature vectors stored in FAISS of desired object
- Cosine similarity: return object class of feature that it most similar if similarity>threshold

Select Personal Item

- Aperol
- Basmati Rice
- Cous cous
- Cous cous2
- Cup
- Rigatoni
- Stelline

Back

Evaluation: Set up

- Cluttered shelf with various objects
- Distance: 5 Feet
- Camera pointed at shelf
- Three products tested (three tries each)
- Products Generic OD:
  - Apple
  - Milk Bottle
  - Beer Bottle
- Products Personal OD:
  - Cous Cous
  - Rigatoni
  - Stelline

Evaluation: Results

	BeMyAI	Meta Glasses	NaviSense	**SeeEverything: Generic Objects**	**SeeEverything: Personal Object**
Search Time (s)	48.23	21.45	15.86	**0.39**	**10.52**
Guidance Time (s)	15.97	19.35	15.89	**7.75**	**1.5-2**
Total Time (s)	64.19	40.8	31.75	**8.15**	**12.5**
Undesired Objects	2.65	4.37	0.52	**0.78**	**0**
Accuracy (%)	85.71	55.88	95.37	**56.25**	**100**

Note:
Undesired Objects in our case defined as Wrong objects identified+ Tracker shift away from object
Accuracy defined as Succesful Retrivals / Total trackings

Comparison to other systems

	SeeEverything	NaviSense	WorldScribe	ThirdEye	MetaGlasses	Recog	Be My Eyes
Open World Detection	✅	✅	✅	❌	✅	❌	✅
Platform independent	✅	❌	✅	Not specified	❌	❌	❌
Personal Objects	✅	❌	❌	❌	❌	✅	✅
No equipment needed	✅	✅	✅	❌	❌	✅	Human volunteer needed
Provides Guidance	✅	✅	❌	✅	❌	❌	✅

Future Work

- Warning when obstacles appear
- Guidance in terms of directions to keep object in frame
- Precise Hand Guidance
- Absolute distance measurements

Conclusion

- Assistive systems for everyone, everywhere
- Combination of Open World OD + Personal Scanning pipeline to optimize finding all possible types of objects
- Real-time detection and navigation

Thank you for your attention!

References

- https://visionatlas.iapb.org
- https://arxiv.org/html/2401.17270v2
- https://dl.acm.org/doi/epdf/10.1145/3313831.3376143
- https://worldscribe.org/

Image References

(1) Sridhar, A. N., Qiao, F., Troncoso Aldas, N. D., Shi, Y., Mahdavi, M., Itti, L., & Narayanan, V. (2026). NaviSense: A Multimodal Assistive Mobile Application for Object Retrieval by Persons with Visual Impairment (Preprint).

(2) https://images.pexels.com/photos/268862/pexels-photo 268862.jpeg?w=940&h=650&auto=compress&cs=tinysrgb

(3) https://www.istockphoto.com/it/immagine/eye-world