MemoVision: A Digital Catalog for Everyday Interactions

Lai Xing Ng, Keith Tien Wei Tang, Jacky Jie Wei Tan
Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR)
{ng_lai_xing;keith_tang;jacky_tan}@a-star.edu.sg

Abstract

We present MemoVision, a digital catalog system that captures *semantic*, *spatial*, *temporal* and *interaction* information as users move around physical environments using client devices such as smart glasses. The system utilizes open-vocabulary semantic segmentation and 3D scans to store objects-of-interest with comprehensive *semantic*, *spatial*, *temporal* and *interaction* labels. Our demonstration shows multimodal information query and retrieval capabilities, supporting specific queries about object locations, temporal events and user interactions including eye gaze and hand poses, enabling more contextualized responses compared to current multimodal large language models.

Introduction

In recent years, advances in multimodal (vision-audio-text) large language models and efficient AI have enabled intelligent AI assistants that can response in real-time to queries from videos, audio and text. Notable technological demonstrations from Google Project ASTRA (Google DeepMind, 2024) and Meta Orion AR (Meta, 2024) showcased advanced capabilities such as agentic workflow to handle and perform tasks and contextual memory that remembers past queries. A user can capture an image or share a live camera with an AI assistant and query it with text or speech. These multimodal inputs are tokenized, encoded and projected into a common embedding space that capture deeper and more nuanced representations across different modalities. Responses are generated using large language models (LLMs).

While the multimodal embeddings can provide a rich representation of the semantics in a scene, spatial and temporal relationships between objects are not encoded precisely. For example, in (Google YouTube Channel, 2024), the user queries where the glasses are and the response is "Your glasses is next to an apple". This information does not provide the exact spatial location and is only useful to the user if the user remembers where the apple is. Therefore, there is a need to provide more specific spatial information with respect to the user's location.

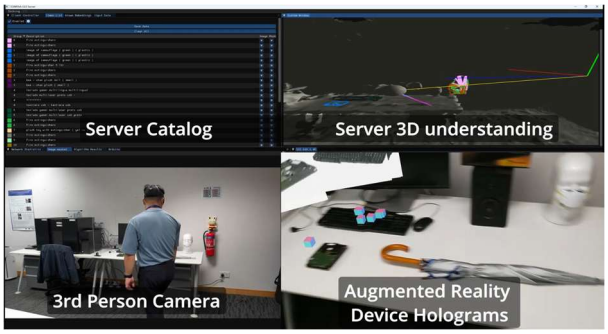

Figure 1: Overview of the MemoVision system (Top: Digital catalog and 3D scene with bounding volumes; Bottom: 3rd person view of a user and client device display).

Current AI assistants only capture scene information through a camera. User interactions with the environment such as eye gaze and hand interactions, are important information to provide personalized context. Objects that a user interacted with are likely to be queried and current systems do not capture how, where and when interactions happen.

To address these limitations, we propose MemoVision, (Figure 1), which is a digital catalog that is built when a user is moving around a physical environment, storing semantic, spatial, temporal and interaction information that can be retrieved. Our contributions include:

- A digital catalog build-up system that captures images from multiple viewpoints, 3D point cloud, meshes, user interactions and semantic embeddings for object-of-interests to provide more holistic 3D scene understanding.
- A multimodal information query and retrieval pipeline that support more specific queries with spatial, temporal and interaction augmenting semantic information.

MemoVision System

MemoVision is a digital catalog (Figure 1) that is built up by users navigating and interacting with a physical environment to capture scene information via smart glasses and mobile phones. As a user moves around, MemoVision utilizes

open-vocabulary semantic segmentation and simultaneous localization and mapping (SLAM) algorithms and 3D scans to store a catalog of identified objects with *semantic*, *spatial*, *temporal* and *interaction* information. The digital catalog can be queried using multimodal inputs (image and text) and information are searched across the four types to generate responses on how, where and when the user has interacted with objects-of-interest.

Digital Catalog Build-Up

MemoVision employs a client-server architecture to build up the digital catalog, with the client device providing multiple data streams (camera, depth, camera pose, eye gaze and hand poses) to the server, which performs computational-heavy processing to (1) fuse the sensor data, (2) identify and segment objects-of-interest and (3) label 3D bounding volumes in the 3D map.

Data Fusion: The data streams from the client devices have different sampling rates and are transmitted to the server at different times. A data interpolation algorithm is developed to handle data losses and incomplete data in between time frames so that the system can access most data types at all time frames. For a specific time frame, the sensor data from the nearest pre and post time frames are used for interpolation. Derived data from the client includes 3D scans and user poses (head, eye gaze and hands) and the system stores the last-known information as client device transmits default data when there is no update.

Segmentation of Object-of-Interest: MemoVision is designed to support open-set object detection and semantic segmentation to handle and store unknown objects. With the input of a single RGB image, the system finds object proposals (up to 10 proposals) using an object detection algorithm (Ren et al., 2015), and for each proposal, a mask is generated using a class-agnostic semantic segmentation algorithm, (Kirillov et al., 2023). The masked patches are also used to generate image-text embeddings using CLIP (Cherti et al., 2023; Open CLIP, 2025) as *semantic* labels in the digital catalog.

Generation of 3D Bounding Volumes: The time-synchronized masked patches and camera poses are placed in the 3D map and ray-casting is performed to find the 3D points and meshes that correspond to the mask. A 3D bounding volume algorithm is used to group the 3D points together as a bounding volume and refine them as additional data is captured from different viewpoints and time frames, i.e. multiple mask patches are used for each 3D bounding volume. The 3D volumes of object-of-interests are *spatial* labels, storing their locations and rough sizes.

Capture of User Interactions: With the 3D volumes and map, the system can know when objects-of-interest appear in the scene as well as the type of user interactions via the head, eye gaze and hand poses. Three types of user interactions are supported: (1) "Appeared" when the object is in the user's field of view, (2) "Looked At" when a user's eye gaze is fixated on an object for more than 2 seconds, (3) "Touched" when a user uses the hands to interact with an object. The time of interaction and type of interactions provide the *temporal* and *interaction* labels respectively.

Querying Objects-of-Interest

Users can query for objects-of-interest stored in the digital catalog in MemoVision to retrieve their locations and time of past interactions. The system can support image and text queries by first encoding the inputs to multimodal embeddings and a distance metric such as Cosine Similarity is used to find the most similar *semantic* labels. As open-vocabulary algorithms are used, the queries can be open-ended and include appearance attributes such as color and textures. Depending on the type of queries, the system can return relevant *spatial*, *temporal* and *interaction* information. For example, when the user asks, "Where is my glasses?", MemoVision can provide directional and location guides overlay on the client device so that the user can find the object with respect to his/her current position. In addition, to aid user memory, the system can prompt the user to interact with the object via hand or eye gaze.

MemoVision can support more specific queries that include *spatial*, *temporal* and *interaction* information such as "Is X next to Y?", "What is the largest object in the room?", "When did I do X?", "What has been moved at Time X?", "What have I looked at/touched?". Retrieval is fast and trivial as constraints can be set on the type of information to limit the search space for semantic labels.

Demo Implementation

For the MemoVision demonstration presented in this paper, the following hardware and software are used:

- Client device: HoloLens 2 (Microsoft HoloLens, 2025)
- Server: CPU-AMD Ryzen 9 5950X 16-Core Processor, GPU-Nvidia RTX3090, 64 GB RAM
- Open-source software: UWP (Microsoft, 2023), OpenXR (Khronos Group, 2025) Bullet Physics Library (Bullet Physics, 2024), Quickhull (Barber et al., 1996), FasterRCNN (Ren et al., 2015), Segment Anything Model (Kirillov et al., 2023), OpenCLIP (Open CLIP, 2025).

Discussion and Future Works

The current version of MemoVision is an alpha version with basic functionalities to support *semantic*, *spatial*, *temporal* and *interaction* information retrieval that are less supported by current vision-LLMs and multimodal LLMs. Exact spatial locations, size information, temporal and user interaction events can augment the responses from LLMs to user queries. MemoVision complement existing works in open-vocabulary semantic segmentation (Liu et al., 2024) with additional spatio-temporal and user interaction information.

Some of the future works include integration with existing multimodal LLMs where MemoVision digital catalog can be used as a database for Retrieval-Augmented-Generation (RAG); extension of the current system to support multiple users so that shared memories and spaces can be built; and handling of dynamic objects and status changes including when they are not captured by cameras.

Acknowledgements

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. 1996. "The quickhull algorithm for convex hulls". *ACM Transactions on Mathematical Software*. 22 (4): pp. 469–483. doi:10.1145/235815.235821.

Bullet Physics SDK. 2023. https://github.com/bulletphysics/bullet3. Accessed 01 Aug 2025.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; Jitsev, J. 2023. Reproducible Scaling Laws for Contrastive Language-Image Learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023, pp. 2818-2829, doi: 10.1109/CVPR52729.2023.00276.

Google DeepMind. 2024. Project Astra. https://deepmind.google/models/project-astra/. Accessed 01 Jul 2025.

Google YouTube Channel. (2024). Project Astra: Our vision for the future of AI assistants. https://youtu.be/nXVvvRhiGjI?t=80. Accessed 01 Jul 2025.

Khronos Group. 2025. OpenXR SDK. https://github.com/KhronosGroup/OpenXR-SDK. Accessed 01 Aug 2025.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; and Lo, W.; Dollár, P.; Girshick, R. 2023. Segment Anything. *IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 3992-4003, doi:10.1109/ICCV51070.2023.00371.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; Zhu, J.; Zhang, L. 2024. Grounding DINO: Marrying DINO: Grounded Pre-training for Open-Set Object Detection. *In Computer Vision – ECCV 2024: 18th European Conference*, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVII. Springer-Verlag, Berlin, Heidelberg, pp. 38–55. doi:10.1007/978-3-031-72970-6_3

Meta. 2024. Introducing Orion, Our First True Augmented Reality Glasses. https://about.fb.com/news/2024/09/introducing-orion-our-first-true-augmented-reality-glasses/. Accessed 01 Jul 2025.

Microsoft HoloLens 2. 2025. https://learn.microsoft.com/en-us/hololens/. Accessed 01 Aug 2025.

Universal Windows Platform SDK. https://developer.microsoft.com/en-us/windows/downloads/windows-sdk/

Open CLIP. 2025. https://github.com/mlfoundations/open_clip. Accessed 01 Aug 2025.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. *In Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15)*, Vol. 1. MIT Press, Cambridge, MA, USA, pp. 91–99. doi:10.5555/2969239.2969250.