Study on the Feedback Modes of Memory Augmentation System for Everyday Life

Lukianova Elizaveta, Jin-Woo Jeong
Department of Data Science, Seoul National University of Science and Technology

el.lukianova@seoultech.ac.kr, jinw.jeong@seoultech.ac.kr

Abstract

This paper introduces a Hololens 2-based system designed to enhance human memory through visual feedback. The system captures and processes images of the user's surroundings, storing them with detailed descriptions. Users can retrieve this information later by asking questions, with feedback provided in Visual, Textual, or Combined formats. A case study with three participants showed a preference for the Combined mode, which effectively supports memory recall by offering both images and text. The findings suggest that the system could be a valuable tool for improving memory retrieval in everyday scenarios.

Ⅰ. Introduction

People see lots of things every day, but, since human memory is imperfect, it is hard to remember them all. However, one may need such information later: we may need to remember at some moment where we placed a thing or what was the place we saw like.

Though there have been a lot of works dedicated to human memory augmentation, few works aim at helping one with open problems, i.e. problems freely formulated by a user asking a question rather than a predefined set of tasks and problems. The recent development of Multimodal Large Language Models (MM-LLM) allowed such freedom. As a result, MM-LLM-based systems like Memoro [1] and system in [2] have been proposed. These two systems provide memory-related audio and text feedback to the user; however, for a human, visual information (imagery) can be processed faster than the audio and text leading to better recall and recognition and more vivid memory revitalization. Alas, there has not been much research on imagery as the main source of memory feedback in daily life.

In this work, we design and implement a system providing a user with visual feedback based on previously seen places and objects. We also conduct a case study with three participants to prove that such a system can be well applied in daily situations and visual feedback is preferred by the users over text.

Ⅱ. Proposed Method

To seamlessly incorporate our system into daily life while still providing swift visual feedback, we implement our system on a Hololens 2, a modern Optical Head-Mounted Display (OHMD), with visual feedback as embedded holograms. The main three stages of the process of collecting information and its further processing can be divided into 1) Encoding for capturing the user's field of view and processing of these consecutive captured images, 2) Storage for efficient storage of these processed frames for future use as visual feedback, and 3) Retrieval for finding a certain frame from the collection most accurately answering the question posed by a user and presenting it as a hologram in user's view. We also enable three types of feedback in our system: Visual (retrieved image), Textual (only text based on image and explaining how an image is an answer to a question), and Combined (showing both Image and Text). These three types are all presented to prospective users in our Case study for comparison based on usability and better use as memory-related feedback format. The outline of the system is demonstrated in Fig. 1.

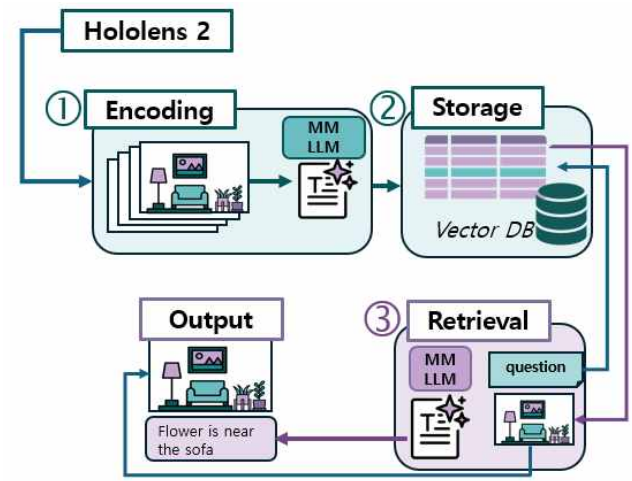

Figure 1. *Outline of the proposed system.*

The main three stages mentioned earlier can be described in detail as follows. 1) Hololens 2 built-in camera captures the user's view and sends images to a Linux server via socket at 2 frames per second. The frames are filtered and only those with variation of Laplacian higher than 20 and less FLANN-based matches with the previously saved image then 100 based on SIFT descriptors are saved; this way, we ensure images are not blurry and not repetitive.

Textual descriptions of frames are generated using one of the best well-known and best-performing MM-LLMs, GPT-4o, with tailored prompts designed so that the MMLLM describes a given frame with as many details as possible ensuring all the properties of objects and their spatial relations. These descriptions are stored in Chroma, a vector-processing database, together with references to the original images previously saved and stored on the server. Texts are vectorized using OpenAI's text-embedding-ada-002 model to allow semantic similarity searches.

When a user submits a text query using voice input, it is vectorized and used to find the most similar frame description in the database via cosine similarity. The matching frame is retrieved, and MM-LLM refines the answer by generating a textual response trying to answer the question using the retrieved image as a reference. Both the frame and its description are sent back to Hololens 2.

The user interacts with the system through a Hololens 2 app built with Unity and the Mixed Reality Toolkit. Feedback is available in three modes: Visual, Textual, and Combined. To ask a question, the user presses a recording button, triggering speech-to-text enabled via Microsoft Azure services. When the question is sent, it is used as a query as previously described. The answer is shown accordingly to a selected mode.

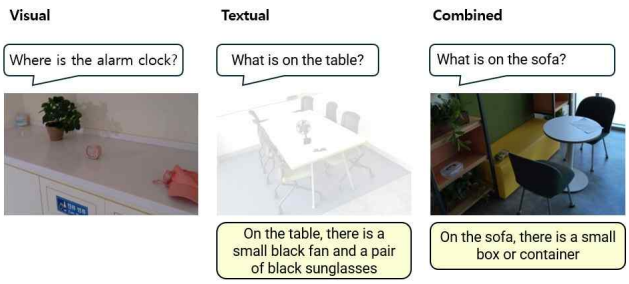

Figure 2. *Examples of the answers generated by the system (in Textual, image is not visible and is shown for reference).*

III. Case Study

In our case study, we invited three participants, local students of age 23-26 years, to test our system and provide feedback on it. Participants wearing Hololens 2 made a tour around the one of the campus buildings for about 10 minutes. Then, after explanation on how to use Hololens 2 and the system, they answered questions asked by interlocutor using any mode of the system they liked. The questions concerned the places and objects previously seen by participants and recorded as frames by the system. Examples of questions and answers are shown in Fig. 2.

All of the participants confirmed they liked the Combined mode the most since it provided both images which allowed fast recognition and text which helped to better understand the image. P1 mentioned that they "looked at image as the main modality and used text as supporting to locate the object in question in the image". Participants suggested that such system could be used to help someone find previously seen things and better explain them to others. P2 even mentioned that such system could be used in specific cases, i.e., communicating some witness evidence to police. However, participants also proposed some directions for improvements. In particular, they voiced the need for fixed in space menu and better and more intuitive user interface. P3 also mentioned blurriness of some frames appearing as answers.

IV. Conclusion

Our study demonstrates the potential of a Hololens 2-based system that leverages visual feedback for memory augmentation. By capturing and processing visual data, storing it efficiently, and retrieving it based on user queries, we created a system that provides versatile feedback in visual, textual, or combined formats. The case study results suggest that users prefer the combined mode, which offers both rapid image recognition and textual explanations, enhancing recall and understanding. This approach shows promise for practical applications in daily life, such as aiding in locating objects or recalling details of previously seen places. Future work could further build on participants' opinions and improve user interface and explore optimizing the system for different user needs and expanding its applicability across various contexts.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (2023R1A2C1006300)

참 고 문 헌

[1] Zulfikar, W. D., Chan, S., & Maes, P. "Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation", Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1-18, May 2024

[2] Shen, J., Dudley, J., & Kristensson, P. O. "Encode-Store-Retrieve: Enhancing Memory Augmentation through Language-Encoded Egocentric Perception", arXiv preprint arXiv:2308.05822, 2023