Voxento-Pro: An Advanced Voice Lifelog Retrieval Interaction for Multimodal Lifelogs

Ahmed Alateeq
ahmed.alateeq2@mail.dcu.ie
School of Computing
Dublin City University
Ireland

Mark Roantree
mark.roantree@dcu.ie
Insight Centre for Data Analytics
Dublin City University
Ireland

Cathal Gurrin
cathal.gurrin@dcu.ie
School of Computing
Dublin City University
Ireland

ABSTRACT

We present an advanced version called Voxento-Pro which is an interactive voice-based lifelog retrieval system. This system has been developed to participate in the seventh ACM Lifelog Search Challenge LSC'24, at ICMR'24 in Thailand. In Voxento-Pro, we introduce a conversational query methodology by utilising OpenAI's Assistant API and employ OpenAI's Whisper technology for state-of-the-art speech recognition and synthesis. This novel version features a more natural interaction mechanism, which enhances the user's experience. In addition, the user interface (UI) was redesigned and introduced a new chat interface and other components. The backend retrieval API was rebuilt with a new technology to support fast and efficient API interactions. Data processing of the lifelog data resulted in about 20% of non-important images being identified and 27% of missing data being filled with Geocoding APIs.

CCS CONCEPTS

• **Human-centered computing** → **Sound-based input / output**; • **Information systems** → **Search interfaces**; • **Computing methodologies** → **Speech recognition**.

KEYWORDS

lifelog; interactive retrieval; voice interaction; conversational search

ACM Reference Format:
Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2024. Voxento-Pro: An Advanced Voice Lifelog Retrieval Interaction for Multimodal Lifelogs. In *The 7th Annual ACM Lifelog Search Challenge (LSC '24), June 10, 2024, Phuket, Thailand*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3643489.3661130

1 INTRODUCTION

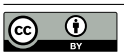
Lifelogging, the practice of digitally documenting one's life through the use of a range of sensors and wearable camera technology, is revolutionising how we capture and recall our daily experiences. This process creates a rich, multimodal dataset consisting of photos, metadata, location data, and biometric metrics data, among others, aiming to provide an integrated view of an individual's daily activities [7]. Yet, the challenge is in our ability to effectively retrieve a particular moment in detail, which is essentially a problem of searching through a vast array of multimodal lifelog data to find relevant content.

Recent development of AI tools and the implementation of large language models (LLMs), provide the ability to understand human natural language through its training on vast amounts of text data. For instance, OpenAI's ChatGPT (Generative Pre-trained Transformer model), employs advanced algorithms to understand and process natural language queries [14]. We believe utilising these advanced technologies can have a significant contribution to pushing the field of lifelog retrieval forward.

Voxento's evolution has resulted in a state-of-the-art lifelog retrieval system that offers a user-friendly voice interaction, facilitating effortless access to lifelog data. This system has been a participant in the last four LSC challenges, showing progressive improvements [2–5]. Hence, we introduce Voxento-Pro which utilises OpenAI's Assistant API [14] to deliver a greater understanding of the context to user queries with improved levels of support when detecting relevant answers from lifelog metadata. We also incorporate the OpenAI Whisper API [16] as an additional speech recognition feature, leveraging its exceptional accuracy and capacity to operate in a variety of noisy environments. In addition, significant effort was made in processing the lifelog dataset, where approx. 20% of the images were classified as not within the user interests scope (not likely to be of relevance to any expected user queries). We retained the OpenAI CLIP model (ViT-L/14) [15] while evaluating larger models from OpenCLIP models (ViT-H/14) and (ViT-g/14) which contain 5 times more trained data [6]. The interface was redesigned with new components such as a chat interface, search-level option and topic task selection. Also, an efficient filtering mechanism was introduced into the system's backend, designed to extract various filters from the query. This approach aims to deliver a focused and relevant set of results, thereby reducing the user's need to manually select filters on the frontend. Finally, this paper includes an analysis of the system's performance evaluation.

2 RELATED WORK

In the recent LSC'23 competition [9], 13 systems participated physically, including the involvement of novice users. LifeXplore, which won for the first time, redesigned its entire system, integrating free text-search using embeddings and utilizing the OpenCLIP (ViT-H/14) model, achieving best results [17]. MyEachtra [20], ranked second, built upon the successful MyScéal system [21], emphasizing event segmentation as a primary contribution to the retrieval process. Both LifeXplore and MyEachtra used OpenCLIP (ViT-H/14) for image-text retrieval. Memento 3.0 [1] also employed CLIP models, leveraging embeddings from a range of larger models like OpenAI



LSC '24, June 10, 2024, Phuket, Thailand

ACM ISBN 979-8-4007-0550-2/24/06
https://doi.org/10.1145/3643489.3661130

CLIP and OpenCLIP, while E-LifeSeeker [12] used similar models alongside CoCa, BLIP, and ALIGN. Overall, there is a growing interest in the use of OpenAI CLIP, OpenCLIP, and BLIP models, which collectively advance the state-of-the-art in lifelog retrieval systems.

In terms of Question-Answering (QA) topics, such efforts made by MyEachtra and MyScéal [20, 21], have laid the groundwork by employing QA models on lifelog data to provide visually correlated answers to queries. Meanwhile, initiatives like E-LifeSeeker, LifeInsight and MemoriEase [12, 13, 23] have utilised embedding models like CLIP and BLIP, though users are required to visually identify answers within images. Our methodology seeks to employ advanced Large Language Models (LLMs) to deliver comprehensive natural language responses to queries alongside the top-ranked relevant images from image-text embeddings, thereby allowing for a more dynamic and engaging user dialogue. This aligns with the essential aspects of question-answering and conversational retrieval, crucial for an efficient voice interface.

Recent advancements in lifelog retrieval are demonstrated by the Lifelog Discovery Assistant [10], which leverages GPT-3 to refine search queries for optimal alignment with vector embeddings through precise instructions. A related work, LifeInsight [13], leverages AI-driven technologies to reformulate user queries into various descriptions relevant to the user's requirements, and is capable of executing multiple retrieval operations using these diverse descriptions. Diverging from this, our method employs GPT-4, catering to the complexity of instructions and system configuration necessary for a conversational chatbot specifically designed to lifelog data retrieval. Our approach enhances query context understanding, provides clarifications as needed, supports speech-based queries, effectively reformulates searches and provides comprehensive answers within the lifelog data.

A few systems incorporate speech-to-text technology to convert speech queries into text, such as [19] by using DeepSpeech technology, yet the discussion on fully voice-based lifelog retrieval systems has been limited, with our system being an exception. Some previous research like [18] has adopted the Speech API following its integration into our work, but none have yet utilised OpenAI's Whisper API. In pioneering this approach, we have introduced Whisper into our system, setting a new standard for accuracy and efficiency in voice-based lifelog retrieval.

3 LSC23 REVIEW OF VOXENTO 4.0 PERFORMANCE

At each participation for Voxento, our highlighted contribution related to system performance, user experience, and new topic queries involving different levels of complexity. Most notably, at the recent LSC'23 competition [9], Voxento 4.0 was ranked 6th of the 14 systems. One interesting observation is that the Novice user performed better than the expert user in all topics of tasks. In addition, the novice user achieved the top score among all participants in the KIS topic. This can be linked to the effectiveness of the interface for novice users. We highlighted the aspects related to Voxento 4.0 performance as follows:

- **User Interface (UI)**: Although we designed the user interface to be more effective, the novice user encountered some challenges at the training session that might be related to the many options and components in the interface. Therefore, minimising the number of components and making the user interface more simple.
- **Ad-hoc topic task**: Despite Voxento's 4.0 efficiency in retrieving many images, a notable portion of the ranked results still includes irrelevant images to the query. To address this, the initial step involves data processing aimed at reducing the presence of non-valuable images, such as blurry ones, from the dataset and eliminating unrelated images through the introduction of a new filtering mechanism.
- **Q&A topic task**: The challenge lies in locating the answer within either the images themselves or the accompanying text in the metadata. Users must carefully examine the images, subsequent events, and relevant metadata to derive an answer. To facilitate the resolution of such complex topic tasks, we leverage the capabilities of OpenAI's Assistant API [14] which employs the API as a retrieval task based on the ranked results set which is derived from the image-text embeddings.

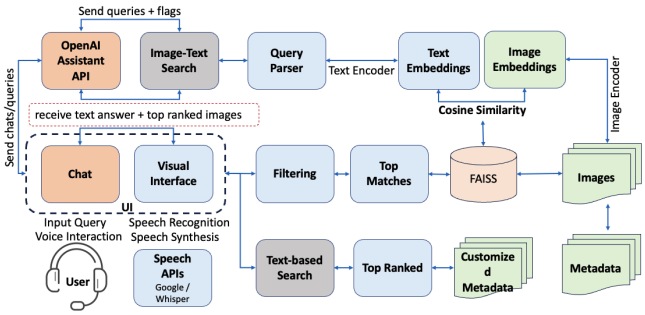


Figure 1: Voxento-Pro System Architecture and Workflow

4 OVERVIEW OF VOXENTO-PRO

In this section, we present an overview of Voxento-Pro and its architecture, with a detailed description of the main components, including the LSC data processing, conversational query methodology, user interface and interaction and semantic search. The system's architecture, illustrated in Figure 1, is divided into five principal components: the user interaction, the visual interface for displaying results, Speech APIs that facilitate voice recognition and synthesis, OpenAI's Assistant API for natural language processing, and the backend API. The backend consists of a filtering mechanism that interprets and implements query filters on search results, along with two search engines—one for text-based queries and another for image-embedding searches.

There is no change to the LSC dataset used in LSC'24 from last year. A detailed description of the LSC'24 dataset can be found in [8, 9]. This dataset contains about 725,000 images which were generated using wearable cameras by an active lifelogger over 18 months between 2019 and 2020.

4.1 Data Processing

As the lifelog dataset expands in size, a key challenge emerges in managing these datasets to align with computational capacities while employing embedding models that maintain a manageable

size of embedding features. To further enhance the dataset, we have undertaken several tasks, including:

- **Non-valuable Images**: We define non-valuable images as those lacking clear content, like blurry or overly dark pictures. Utilising the OpenCV[1] library, we apply a Laplacian variance threshold method for blur detection and a colour analysis strategy for identifying dominant colours in images through HSV (Hue, Saturation, Value) range. This approach helped us identify around 4,000 blurred and 9,000 colour-dominant images, such as those with 9,000 instances of high dominant colour over an image, making them non-valuable. Figure 2 shows examples of such images.
- **Non-important Images**: We assume that such images are likely to be of no interest to lifeloggers seeking meaningful content. Numerous images were discovered that are unlikely to provide any significant meaning, such as pictures of blank walls or ceilings. Our exploration of the OpenCV library revealed that edge detection, specifically through the use of the canny method at a very low threshold, could effectively identify these unwanted images. Approximately 178,000 images, constituting 24% of the dataset, were detected using this method. Figure 2 shows examples of images generated by edge detection.
- **Geographic Coordinates**: We encountered numerous images lacking specific location details, such as city, country, and address. Just having the name of a place proved insufficient for extracting the necessary information. To address this issue, we turned to the OpenStreetMap[2] API and the Google Geocoding[3] API. These tools allowed us to transform the geographic coordinates associated with each image into detailed address information.
- **Travel Labels**: We have decided to further refine and enhance this feature by continuously labelling travel-related images with the format *(source airport code - destination airport code)* from a previous system [5]. We identified the absence of labels for four travel instances. This discovery was made possible by leveraging the VAISL dataset, by [22]. As a result, this enhancement significantly aids in the retrieval process by allowing the system to exclude a total of 17,000 images associated with travel labels when the search query does not pertain to airplanes, thereby streamlining the search experience and improving the relevance of retrieved results.
- **Gaps and Missed Data**: To cover for interruptions in GPS data recording, we implemented a method where we filled in missing location details for images by analysing the temporal and spatial timing of events surrounding the gaps. If events before and after a gap showed the same location and were within a five-minute interval, we assumed the lifelogger remained in the same place, resulting in updating around 200,000 records accordingly. Additionally, we utilised the VAISL dataset [22] to enhance and enrich our metadata with missing information, such as *home* labels, and introduced new types of data like *activity* labels.

4.2 Conversational Query Methodology

It is crucial for LSC to develop a system that supports understanding while providing natural language queries with the interaction in the context of lifelog retrieval. Therefore, we employed the OpenAI Assistant API to guide the search engine to the most appropriate interaction with the user. Hence, by utilising the API we treat the retrieval process as a cooperative task, where the user and the AI API work together to find the desired information.

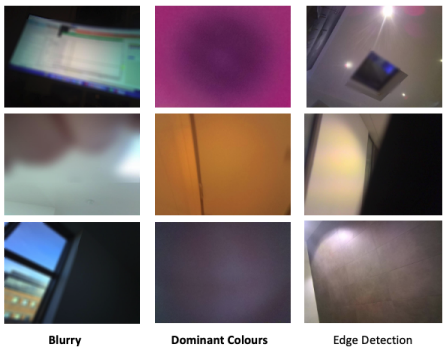


Figure 2: Examples of Non-valuable and Non-important Images

To fulfil the integration smoothly between the user, backend (search engine) and the OpenAI Assistant API, we defined six flags for different contexts of interaction after various designs and tests as follows:

We asked OpenAI's Assistant API in the instructions to include these flags in the response, so the backend retrieval will understand the next action.

- **Reply**: Response for normal query like greetings.
- **Context**: Employ if a query is unrelated to the lifelog data.
- **Clarification**: Required for vague or incomplete queries.
- **Answer**: Submit the answer found in the text.
- **Search**: Appropriate when a query is detailed enough
- **Re-search**: Reformulate the query if the results are insufficient.

Figure 3 displays a variety of chat dialogues between users and the system, facilitated by OpenAI's Assistant API. Furthermore, OpenAI's Assistant API incorporates a retrieval task feature that allows for dataset attachment, yet the full lifelog dataset remains securely stored in our backend and is not fully disclosed to OpenAI. Only a limited selection of the top 10th-ranked list, with specific columns, is shared to ensure privacy considerations are properly upheld.

4.3 Improved Voice-based Interaction

In the context of enhanced voice interaction, the development of the chat interface has been a pivotal step towards achieving a fully interactive lifelog retrieval system. This interface serves as the foundation for voice-based interaction retrieval, transitioning from a traditional standard retrieval system to a lightweight application developed and equipped with a screen capable of processing speech queries and delivering results in both text and image formats.

In terms of new features, we employ state-of-the-art speech recognition and synthesis from the OpenAI API called Whisper [16]. The incorporated additional speech API benefits lifelog retrieval tasks by supporting a more seamless and intuitive user

[1] https://opencv.org/
[2] https://www.openstreetmap.org/
[3] https://developers.google.com/maps/documentation/geocoding/overview

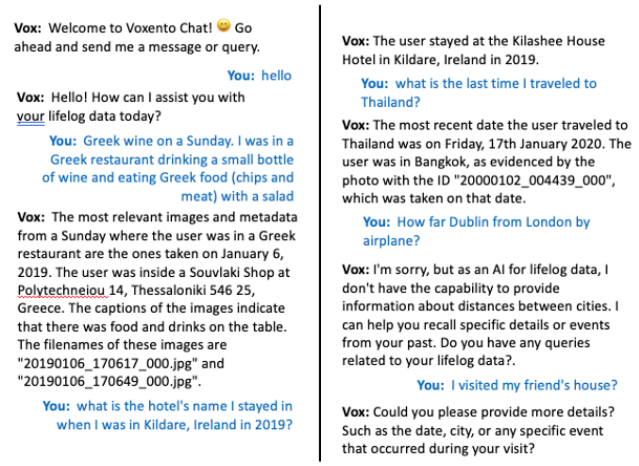


Figure 3: Chat Dialogue Examples

experience, facilitating direct, voice-activated access to lifelog data. Enhancements in voice interaction have been made by removing non-essential voice commands, such as filter applying, and shifting these functions to the backend, simplifying the user experience. Additionally, the interface now supports text-to-speech functionality for responses, enhancing the naturalness of communication.

We have retained the Google Speech API, our native speech recognition technology, ensuring effective interaction by leveraging both technologies. The detailed implementation of voice interaction features in earlier versions of our system is thoroughly documented in [3, 4], providing an extensive overview of the features and methodologies employed in our approach to voice-based lifelog retrieval.

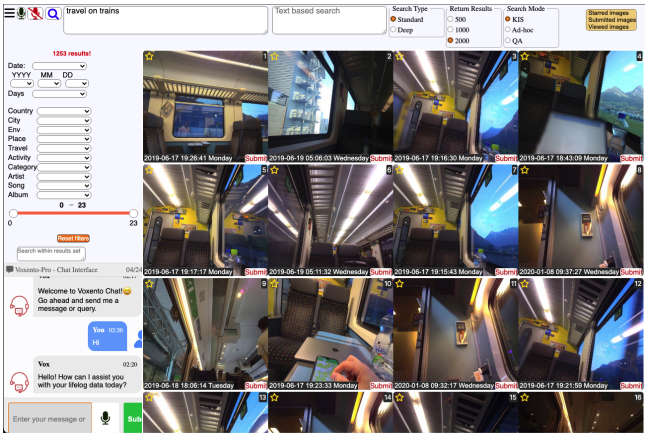


Figure 4: Voxento-Pro Main Interface

4.4 Improved User Interface (UI)

Reflecting on the insights gained from the analysis in section 3, we undertook a comprehensive redesign of our interface to prioritise simplicity and enhance usability for novice users, as shown in Figure 4. The updated user interface incorporates several key enhancements to improve user interaction and system functionality, with retained crucial features from previous versions.

Firstly, We introduced a chat interface that connects to the backend, integrating with the OpenAI Assistant API, equipped with a speech recognition button allowing direct speech-to-text input. We now offer two different levels of search functionality. The optimised search utilises extracted filters and excludes non-valuable images. The second level 'deep search' mode includes all images with filtering, providing a more extensive search option. Additionally, users will have the option to select the desired number of search results, enhancing customization based on user preference. To further assist users, especially novices, we introduced selectable topic types to guide the search process more effectively.

In terms of visual enhancements, the interface now displays a list of activities before and after a selected image and includes detailed annotations within zoomed images. We have also enriched the image labels in the main interface by adding location details alongside existing date and time information. Newly implemented is a method for extracting the frequency of words in the results set, which provides insights that are particularly useful in responding to queries about trends, such as *"Which airline did I fly with most often in 2019?"*.

4.5 Improved Semantic Search

Firstly, regarding the backend API development, we transitioned from using Flask API to FastAPI for our new API technology. Unlike Flask, which operates on a synchronous model, FastAPI functions asynchronously, making it more adept at handling numerous requests simultaneously. This adaptation is especially beneficial given the competitive nature of the LSC, where timely and efficient handling of multiple queries is crucial. FastAPI enhances the speed of our search query API calls, allowing for quicker responses.

In exploring semantic search, the OpenCLIP ViT-H/14 model demonstrated superior performance in LifeXplore and MyEachtra winning systems [17, 20]. Meanwhile, the OpenCLIP ViT-g/14 model showed enhanced performance compared to the OpenAI ViT-L/14 model at hint 3 [1]. However, at the initial hints, the OpenAI ViT-L/14 model outperformed all competing models [1]. Given the proven effectiveness of the OpenAI CLIP ViT-L/14 model in handling most LSC queries, we have continued its use while also evaluating the OpenCLIP ViT-H/14 and ViT-g/14 models.

An advancement was made in optimising text-based search functionality. While text-based search mechanisms were operational in prior versions of our system, the principal challenge lay in the wait time for generating ranked results. To address this, we carefully selected metadata to encompass only essential texts and streamlined the search algorithm to operate on a select subset of metadata columns.

4.6 Improved Filtering Mechanism

In an effort to further improve the efficiency of data retrieval, we innovated a pre-result filtering system. This enhanced mechanism utilises a set of pre-configured filters, capable of being called directly on the query prior to the generation of result sets. These filters consist of detailed criteria, including *date, time,* and *location,*

for instance, within the time filter, specifications such as midnight, morning, afternoon, and even seasonal categories are considered. Moreover, the refinement of our filtering methodology extends to the accommodation of spelling errors within queries related to filter criteria. For instance, a search query such as *I visited Thiland in the last sumer for a university trip*, despite misspellings, is interpreted accurately by the system to deduce intended filters—country name and seasonal period. In this scenario, the system's enhanced linguistic flexibility recognises variations of 'Thailand' as *thiland, thilnd, tailend* and 'Summer' as *sumer, summr*, thereby maintaining the efficacy of the search mechanism despite input inaccuracies.

4.7 Improved Vector Similarity

Major strategies have been employed to enhance the efficiency of vector similarity computations. The initial approach involved narrowing the dataset by removing 20% of images deemed non-essential, leading to the creation of two separate sets of embeddings based on specially prepared metadata. The first set encompasses embeddings for the entire image collection, while the second excludes those identified as non-important. This distinction has facilitated the introduction of a search-level feature within the user interface, empowering users to select the depth of search according to their needs when initial results are deemed inadequate.

In terms of indexing the embedding features, the system leverages the same as the previous version, FAISS (Facebook AI Similarity Search) [11] with a brute-force approach. LifeXplore and Memento 3.0 [1, 17] showed competitive systems using the FAISS and the method of the inverted file index with clustering. In our approach, we have expanded upon this methodology by increasing the number of clusters to 100 through an inverted index strategy. This adjustment significantly accelerated search speeds, however, initial testing found that a brute-force approach could have more accuracy.

5 SYSTEM EVALUATION

Based on the innovations introduced into the current system, we evaluated 24 Known-Item Search (KIS) topics, incorporating 14 from LSC'22 and 10 from LSC'23, utilising the HiT@K metric, which measures the effectiveness of retrieving at least one relevant item within the top-K of the result set. This evaluation employs a range of K values (1, 3, 5, 10, 20, 50, and 100) and the provided information hints at distinct time intervals (T=0,30,60 seconds), which closely follows the Memento evaluation [1]. The evaluation was performed on three versions: $Voxento4.0$, our previous system; $Voxento-ProOpenAICLIP$ the current version; and $Voxento-ProOpenCLIP$, using the OpenCLIP model (ViT-H/14).

The data presented in Table 1 demonstrates the enhanced capabilities of Voxento-Pro OpenAI CLIP ViT-L/14, evidencing its superior efficacy in accurately identifying a specific image from the initially provided hints. Comparing Voxento 4.0 with Voxento-Pro (both versions using the OpenAI CLIP ViT-L/14 model), there's a clear improvement in the HiT@K metrics across almost all K values. The most notable advances are seen in the lower K values (@1, @3, @5), which are critical for immediate user satisfaction as they indicate a higher percentage of the user finding a relevant result quickly. For instance, at hint 1, there's an improvement from 16.67% to 20.83% at @1, and similar incremental improvements are observed at @3 and @5. The user will be able to solve on average 41.67% of the queries in only the top three ranked results with only the first hint. It is important to note, however, that the inclusion of images deemed unimportant marginally impacts the overall result quality. The advancements in data processing have indeed resulted in the successful attainment of the system's performance benchmarks.

Voxento-Pro maintains or slightly improves performance at higher K values, suggesting enhanced overall recall without affecting the precision in early results. In comparison, Voxento-Pro using the OpenCLIP ViT-H/14 model shows a decrease in performance across the board compared to the OpenAI CLIP ViT-L/14 models, even with its expansive training dataset and large model dimensions. This suggests that while Voxento-Pro has benefited from the advanced features introduced in the paper, the OpenCLIP model may not align as well with this particular dataset or the configurations may need further adjustment. Consequently, the employment of both OpenCLIP ViT-H/14 and OpenCLIP ViT-g/14 (also tested but not included in the table) will be excluded from the current version, reserving instead for subsequent testing experiments. The experimental phase shed light on an interesting observation: an increment in hints does not necessarily correlate with a higher probability of retrieving relevant images. The testing analysis shows a slight decrease in HiT@K metrics after the fifth hint.

In our focused evaluation of Question-Answering (Q&A) queries, we selected 10 Q&A topics from LSC'23, on purpose omitting Q&A topics from LSC'22 due to the absence of factual answers. This strategic choice facilitated the examination of the system's efficacy in leveraging the integration of text-image retrieval capabilities with the Assistant API. Among the chosen topics, the system successfully provided detailed responses to 5, offered closely related or relevant answers to 3, and was unable to address 2 topics. Notably, some queries returned responses only after the initial questions were rephrased. The following is an example of a topic for which the system generated an answer:

- **Q&A**: I had some Strawberry Jam / Preserve in my refrigerator. It was the best jam I ever tasted. What brand was it?
- **Answer**: The brand of the best tasting Strawberry Jam / Preserve in the user's refrigerator is "MRS BRIDGES".

6 CONCLUSIONS AND FUTURE WORK

In this paper, we presented an advanced version of Voxento system generation called Voxento-Pro. This paper introduces the conversational query methodology by utilising OpenAI's Assistant API which shows promising results alongside Whisper API, a state-of-the-art speech recognition and synthesis from OpenAI. The user interface was redesigned with the introduction of the chat interface, search-level options and other components. We kept employing OpenAI CLIP (ViT-L/14) and with the system's features showed better results compared to the previous system, specifically in solving on average 42% of queries in only the top three ranked results with only the first hint. Intensive efforts in data processing resulted in identifying about 20% of non-important images, 200k missing data in addresses filled and 16k images travel labelled. We will continue

	Hint	@1	@3	@5	@10	@20	@50	@100
Voxento 4.0 (OpenAI CLIP ViT-L/14)	H1	16.67	37.5	50.00	50.00	66.67	66.67	83.33
	H2	25.00	50.00	**62.50**	**70.83**	**79.17**	**83.33**	83.33
	H3	45.83	50.0	58.33	**70.83**	**75.00**	**83.33**	**83.33**
Voxento-Pro (OpenAI CLIP ViT-L/14)	H1	**20.83**	**41.67**	**54.17**	**58.33**	**70.83**	**70.83**	83.33
	H2	29.17	50.00	58.33	66.67	75.00	79.16	83.33
	H3	**50.00**	**54.17**	58.33	66.67	70.83	70.83	70.83
Voxento-Pro (OpenCLIP ViT-H/14)	H1	16.67	25.00	25.00	33.33	41.66	54.17	62.50
	H2	29.17	37.50	45.83	45.83	45.83	54.17	54.17
	H3	25.00	37.50	41.67	45.83	50.00	58.33	58.33

Table 1: Comparative Evaluation of System Performance for 24 Known-Item Search (KIS) Topics Using Average HiT@K Metrics

improving the conversational query methodology by fine-tuning the model as well as improving the user voice interaction.

ACKNOWLEDGMENTS

This work was supported by Science Foundation Ireland through the Insight Centre for Data Analytics (SFI/12/RC/2289_P2), the Vistamilk SFI Research Centre (SFI/16/RC/3835) and also by the Ministry of Education in Saudi Arabia.

REFERENCES

[1] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2023. Memento 3.0: An Enhanced Lifelog Search Engine for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge (LSC '23)*. ACM, New York, NY, USA, 41–46. https://doi.org/10.1145/3592573.3593103

[2] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2020. Voxento: A Prototype Voice-controlled Interactive Search Engine for Lifelogs. In *Proceedings of the Third Annual Workshop on the Lifelog Search Challenge (LSC'20)* (Dublin, Ireland). ACM, New York, NY, USA, 77–81. https://doi.org/10.1145/3379172.3391728

[3] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2021. Voxento 2.0: A Prototype Voice-controlled Interactive Search Engine for Lifelogs. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan). ACM, New York, NY, USA, 65–70. https://doi.org/10.1145/3463948.3469071

[4] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2022. Voxento 3.0: A Prototype Voice-controlled Interactive Search Engine for Lifelogs. In *Proceedings of the 5th Annual Lifelog Search Challenge (LSC '22)*, Vol. 1. ACM, New York, NY, USA, 43–47. https://doi.org/10.1145/3463948.3469071

[5] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2023. Voxento 4.0: A More Flexible Visualisation and Control for Lifelogs. In *Proceedings of the 6th International Workshop on Lifelog Search Challenge, LSC 2023*. 7–12. https://doi.org/10.1145/3592573.3593097

[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. Reproducible scaling laws for contrastive language-image learning. https://doi.org/10.48550/arXiv.2212.07143 arXiv:2212.07143 [cs].

[7] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. *LifeLogging: Personal big data*. Vol. 8. Now Publishers. 1–125 pages. https://doi.org/10.1561/1500000033

[8] Cathal Gurrin, Liting Zhou, Graham Healy, Werner Bailer, Duc-Tien Dang-Nguyen, Steve Hodges, Björn Þór Jónsson, Jakub Lokoć, Luca Rossetto, Minh-Triet Tran, and Klaus Schöffmann. 2024. Introduction to the Seventh Annual Lifelog Search Challenge, LSC'24. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24)* (Phuket, Thailand) *(ICMR '24)*. ACM, New York, NY, USA. https://doi.org/10.1145/3652583.3658891

[9] Cathal Gurrin, Björn Por Þór Jónsson, Duc-Tien Dang-Nguyen, Graham Healy, Jakub Lokoč, Liting Zhou, Luca Rossetto, Minh-Triet Tran, Wolfgang Hürst, and Klaus Schöffmann. 2023. Introduction to the Sixth Annual Lifelog Search Challenge, LSC'23. In *Proc. International Conference on Multimedia Retrieval (ICMR'23)* (Thessaloniki, Greece) *(ICMR '23)*. ACM, New York, NY, USA.

[10] Nhat Hoang-Xuan, Thang-Long Nguyen-Ho, Cathal Gurrin, and Minh-Triet Tran. 2023. Lifelog Discovery Assistant: Suggesting Prompts and Indexing Event Sequences for FIRST at LSC 2023. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge (LSC '23)*. ACM, New York, NY, USA, 47–52. https://doi.org/10.1145/3592573.3593104

[11] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[12] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Cathal Gurrin, Minh-Triet Tran, Thanh Binh Nguyen, Graham Healy, Annalina Caputo, and Sinead Smyth. 2023. E-LifeSeeker: An Interactive Lifelog Search Engine for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 13–17. https://doi.org/10.1145/3592573.3593098

[13] Tien-Thanh Nguyen-Dang, Xuan-Dang Thai, Gia-Huy Vuong, Van-Son Ho, Minh-Triet Tran, Van-Tu Ninh, Minh-Khoi Pham, Tu-Khiem Le, and Graham Healy. 2023. LifeInsight: An Interactive Lifelog Retrieval System with Comprehensive Spatial Insights and Query Assistance. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge (LSC '23)*. ACM, New York, NY, USA, 59–64. https://doi.org/10.1145/3592573.3593106

[14] OpenAI. 2024. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs].

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 http://arxiv.org/abs/2103.00020

[16] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. (2022). arXiv:2212.04356 https://github.com/openai/http://arxiv.org/abs/2212.04356

[17] Klaus Schoeffmann. 2023. lifeXplore at the Lifelog Search Challenge 2023. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge (LSC '23)*. ACM, New York, NY, USA, 53–58. https://doi.org/10.1145/3592573.3593105

[18] Florian Spiess, Ralph Gasser, Silvan Heller, Luca Rossetto, Loris Sauter, Milan Van Zanten, and Heiko Schuldt. 2021. Exploring Intuitive Lifelog Retrieval and Interaction Modes in Virtual Reality with vitrivr-VR. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan). ACM, New York, NY, USA, 17–22. https://doi.org/10.1145/3463948.3469061

[19] Florian Spiess and Heiko Schuldt. 2022. Multimodal Interactive Lifelog Retrieval with vitrivr-VR. In *Proceedings of the 5th Annual Lifelog Search Challenge (LSC '22)*, Vol. 1. ACM, New York, NY, USA, 38–42. https://doi.org/10.1145/3512729.3533008

[20] Ly Duyen Tran, Binh Nguyen, Liting Zhou, and Cathal Gurrin. 2023. MyEachtra: Event-Based Interactive Lifelog Retrieval System for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 24–29. https://doi.org/10.1145/3592573.3593100

[21] Ly-Duyen Tran, Manh-Duy Nguyen, Binh Nguyen, Hyowon Lee, Liting Zhou, and Cathal Gurrin. 2022. E-Myscéal: Embedding-based Interactive Lifelog Retrieval System for LSC'22. In *Proceedings of the 5th Annual Lifelog Search Challenge (LSC '22)*. ACM, New York, NY, USA, 32–37. https://doi.org/10.1145/3512729.3533012

[22] Ly-Duyen Tran, Dongyun Nie, Liting Zhou, Binh Nguyen, and Cathal Gurrin. 2023. VAISL: Visual-Aware Identification of Semantic Locations in Lifelog. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part II*. Springer-Verlag, Berlin, Heidelberg, 659–670. https://doi.org/10.1007/978-3-031-27818-1_54

[23] Quang-Linh Tran, Ly-Duyen Tran, Binh Nguyen, and Cathal Gurrin. 2023. MemoriEase: An Interactive Lifelog Retrieval System for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 30–35. https://doi.org/10.1145/3592573.3593101