Voxento 4.0: A More Flexible Visualisation and Control for 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

In this paper, we introduce Voxento 4.0 – an interactive voice-based retrieval system for lifelogs which has been developed to participate in the sixth Lifelog Search Challenge LSC'23, at ACM ICMR'23. Voxento has participated three times in the LSC editions and achieved the rank of 4th in LSC21 and 5th in LSC22 respectively. In this version, Voxento 4.0, we have focused on improving the previous system's interface, voice interaction and retrieval functionality. The current version has implemented some processing and cleaning of the dataset and employs the CLIP model to extract image features. In addition, the system's interface was redesigned for better visualisation of the elements and the images for effective interaction. This improvement in the interface will help to support voice interaction in future work. The interface developments include logging voice interaction and images displayed, submitted, selected and starred to enhance user experience with the system. The voice interaction part has also been enhanced in the workflow of the voice lifecycle interaction and with additional voice commands.

CCS CONCEPTS

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

KEYWORDS

lifelog; interactive retrieval; voice interaction; speech recognition; speech synthesis

ACM Reference Format:
Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2023. Voxento 4.0: A More Flexible Visualisation and Control for Lifelogs. In *6th Annual ACM Lifelog Search Challenge (LSC '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3592573.3593097

1 INTRODUCTION

Imagine a human trying to remember a past event that occurred in their life many years ago, such as experiencing a new interesting place or eating exotic foods. We as humans may not be able to remember these events in much detail, or even locate it within our personal photo archives for viewing or sharing. With the development of technologies, applications and sensors, it is possible to have an efficient and lightweight retrieval search engine that can search in a lifelog multimodel record of a human life experience. This process of capturing such information is called lifelogging. Technological examples include wearable cameras, smartphones, and smartwatches [8]. People nowadays are increasingly using their phones for capturing important moments and share them via social applications or keep them in the phone's storage, which can be considered a part of the lifelogging concept. Hence, we still sometimes find it difficult to even find an important image in our phones, although new technologies were developed in the search engine of the phone's gallery such as identifying face images and searching through text inside the image. Therefore, a lifelog can have much more data than typical smartphones, and the multimodal nature of lifelogs makes this an even more challenging task. Because of the diversity of sensors and technologies that are available to gather lifelogs, such lifelog data must be reorganised, restructured and integrated into a specific form to facilitate effective information retrieval.

Since lifelogging as a concept was not yet widespread and multimodal lifelog data was not yet commonly available, the Lifelog Search Challenge (LSC) was founded in 2018 as an annual benchmark competition workshop to identify the state-of-the-art approaches to address the challenge of information retrieval from multimodal personal lifelogs. Over the past six years, the LSC challenge has experienced new changes in terms of evaluation processes, topic types and dataset size and variety. LSC's query tasks became increasingly difficult and realistic in recent years as the sophistication of the state-of-the-art approaches has increased. The most recent LSC'22 challenge introduced two new types of topics in addition to known-item topics. Firstly, Ad-hoc search, which is to find as many relevant images as possible from the lifelog that answer the topic. Secondly, Q&A topics seek a correct answer to an information need and this will be judged in real-time. In terms of the LSC workshop process and the details of how the workshop runs are described in [10].

The aim of Voxento is to be a simple lightweight application that can be deployed for mobile devices or on smart TVs where the user or lifelogger can communicate and interact with a personal lifelog retrieval by a voice as a means of interaction. Voice-controlled lifelog retrieval is not well understood yet and this motivates us to continue developing and improving the system. In this paper, we present the fourth version of Voxento, a prototype voice-controlled interactive retrieval system for lifelogs. Voxento participated three times since LSC'20 [2], LSC'21 [3] and LSC'22 [4]. The system

LSC '23, June 12–15, 2023, Thessaloniki, Greece
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0188-7/23/06.
https://doi.org/10.1145/3592573.3593097

ranked 4th in LSC21 (top-performing systems, scored 91 out of 100) and 5th in LSC22. In each edition of LSC, there were some challenges and issues discussed in the next published paper of each system and proposed proper suggestions on the improvements.

In the system presented here, we have focused on enhancing the system interface, dataset restructuring and the backend retrieval functionality, based on the experiences from the previous LSC challenges. The main interface has been redesigned and redeveloped for better structuring and distributing of the elements to have filter selection on one side and query text bar on the other side. This will create more flexible visualisation and user control. We also worked on analysing the dataset to see where data cleaning or enhancement can be made. For the backend, we also utilise the CLIP model [16] to extract features from images in order to have high accuracy when comparing the similarity between the text features with the image features. A further contribution of this research includes sharing the dataset from analysing the location of airplane travel for usage by the LSC organisers.

2 RELATED WORK

Since the development of the first lifelog retrieval system in 2006, called MyLifeBits [6], there are in excess of 70 lifelog retrieval systems that have been developed until today, including systems with different versions that participated multiple times in different benchmarking challenges. These systems were mostly listed and categorised into main focuses to understand trends in lifelog development. Most systems saw much of the effort spent on metadata enhancement. Consideration of how users interact with the retrieval system has received less attention from researchers and there are only a few systems that support non-desktop-based systems. Examples of systems focused on the interaction are ViRMA [5] in Virtual Reality and XQC [14] for mobile devices. This is one of the motivations of this research: to focus on developing a new way of interaction to support the easy accessibility to the lifelog data, in our case, via voice.

In the most recent LSC'22 competition, 9 systems participated [9]. The system which performed best was E-Myscéal 2.0 [19] and it has been the winner since the LSC'20 edition. E-Myscéal 2.0 employs an embedding technique by the CLIP model that facilitates novice users who are not familiar with the concepts. Notably, Memento 2.0 [1] also used the CLIP model by leveraging the embeddings generated using two larger CLIP models (ResNet-50) and (ViT-L14). Our previous system, Voxento 3.0 [4] used the latest and a larger model called (ViT-L14@336px) which was pre-trained on a higher pixel resolution for better performance. In addition, Lifeseeker 4.0 also employs the CLIP model to improve retrieval functionality. Another system that participated is Vitrivr [11] which supports text embedding, text retrieval, Visual similarity and boolean retrieval. Vitrivr developers have enhanced a second version of the system that uses Virtual Reality [18]. LifeXplore, [15], a system that participated since the first LSC'18 [7], has enhanced the current version of the results presentation and search interface. The system supports viewing images based on calendar filters in chronological order. The last system is FIRST 3.0 [12] also utilised the CLIP model for image-text embedding.

An additional observation is that when the lifelog dataset gets larger, there have been attempts to reduce the size of the dataset while also maintaining enough semantically valuable content to support effective retrieval functionality. Although the contribution of our system Voxento is in the voice interaction, we keep enhancing and focusing on all aspects of an efficient search engine including improving interface design, metadata, extracting image features and search engine ranking and indexing. Based on our experience with the LSC competition, we believe that metadata pre-processing, cleaning and restructuring is a crucial factor in affecting the system performance.

3 VOXENTO'S PERFORMANCE AT LSC 2022

Voxento 3.0 was ranked 5th overall at LSC'22 [4], similar to the previous ranking at LCS'21. One issue that affected the performance of the system is that some incorrect answers had been submitted for some queries, due to misunderstanding the information needs and based on the nature of the competition which required a fast submission and searching through the images. In each participation, logging of actions provided important feedback that could be used to improve the user interaction and retrieval tasks. The first note is that the system was not implemented to answer the new search topics more efficiently. For instance, submitting a number of relevant images was limited by the fact that the system required manual submission of each image, which made the submission of multiple valid images quite slow. The second important note is that the distribution of the elements and the search filters in the interface did not support user interaction effectively. We have redesigned the entire interface from scratch. The configuration of the voice interaction options was moved to its own window. The search filters moved to another space, ordered horizontally. The last note is regarding the utilising of the CLIP model for extracting image features. Since retrieval performance was impacted by the volume of embedding features, we spent time pre-processing the dataset to reduce the index size and therefore improve retrieval speed. In addition, the voice interaction was limited to certain aspects of the interface. Thus, we need to expand the coverage and usefulness of the voice commands.

4 LSC'23 DATASET

LSC'23 uses the same challenge dataset as the previous year. We now highlight the aspects of the LSC'23 dataset in this section, with a more detailed description available in [10]. The LSC'23 dataset was generated by one active lifelogger over an 18 month period during the years 2019 and 2020. The number of images captured is approximately 725k images. The full dataset included fully anonymised and redacted images, visual concepts, metadata, and text captions. Visual concepts include text descriptions of detected scenes, concepts, and objects for each image with the confidence score. The metadata contains time, date, physical activities such as *steps*, biometrics such as *heart beats*, *calories*, *sleep efficiency*, locations such as *country name*, *geographic location*, music information such as *artist, songs, album*, and other metadata.

Based on the experience in participation in LSC and developing a lifelog retrieval system, there are main tasks in processing the dataset. Firstly, deriving features from the metadata and separating

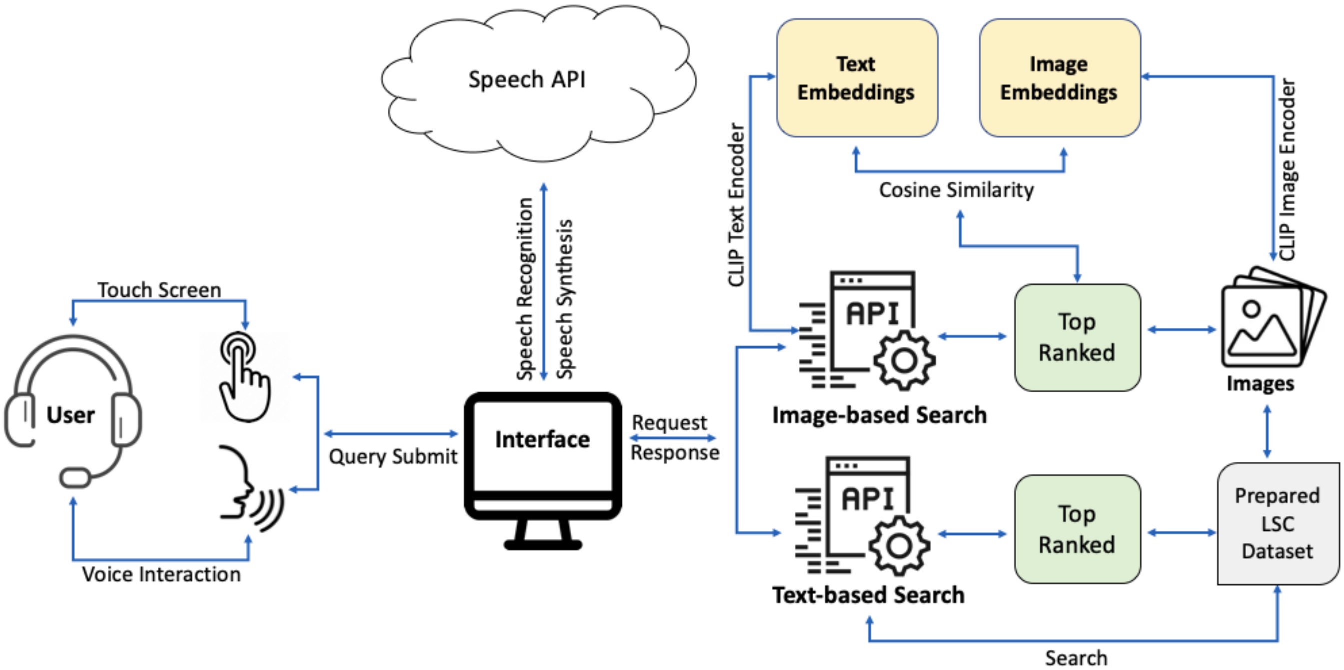

Figure 1: Overview of the System Architecture and Workflow

the columns such as the date and time (day, month, year, hour, and minute). Following this is to extract valuable information to be used as filters and enhance the retrieval such as extracting such as day and month names that were represented as numbers. In addition, in regard to the cleaning dataset, we used the same method in the previous tasks including the use of the CLIP model by attempting to identify blurred images and meaningless images such as dark or all-black background images. These images might have been taken by accident when the lifelogger was in a dark room or obscured the camera lens (e.g. by clothing or fast-moving hands).

In the previous system, we incorporated steps to reorganise and restructure the data to remove some artifacts of the automatic data gathering. One of the tasks completed for LSC'23 was solving one of the issues with automatic location logging. When GPS logs are recorded continually and the wearer is in motion (car, airplane), the location of the lifelogger will follow the travel route, though they may never actually have been active in these locations. So, when the lifelogger was flying, these location entries were replaced with the source and target locations (e.g. when travelling from Dublin to Istanbul, the location is replaced by DUB-IST. This task is mostly automated though it requires some manual validation. This list of airplane travelling locations will be made available to all participants. Another task is adding the city name to all locations, which is not always provided by default.

5 OVERVIEW OF VOXENTO 4.0

In this section, we present an overview of the Voxento 4.0 system and architecture, with a detailed description of the main components. Voxento 4.0 integrates a number of enhancements, namely the interface design, the search engine and the voice interaction. The system architecture and the workflow presented here are similar to the previous system but we illustrate an updated system architecture, providing further detail about the system components.

The architecture and the workflow of the system can be viewed as having four main components as shown in Figure 1: the user's voice and interface interaction, the visual interface and the Speech API, and the backend API that has two search engines. One search engine for the image-based search uses embeddings leveraged from the images by the CLIP model. Another is the text-based search engine for searching through the text in specific columns in the prepared and pre-processed LSC dataset from the LSC metadata associated with each image. Both retrieval APIs are linked to the dataset of images located in the server.

5.1 Improved Voice-based Interaction

In the voice interaction functionality, we continue with the same methodology and implementation, but while this appears similar to previous versions, there is continuous improvement to Voxento's services for a better user experience. For details as to how the voice interaction is implemented and the way users interact with the system, refer to [3] and [4]. Briefly, the Voxento interface was developed as a web-based application and the user uses a headset for voice interaction or a mouse and keyboard for standard interaction support if needed. The user can swap between voice interaction and using a mouse and keyboard at any stage. The system is configured to catch specific voice commands at various stages of the retrieval process, such as when the system accepts the selection of image commands if there are images displayed and the query has already been submitted to the server. The speech API is provided by the Google Web Speech API through the Chrome browser. We also

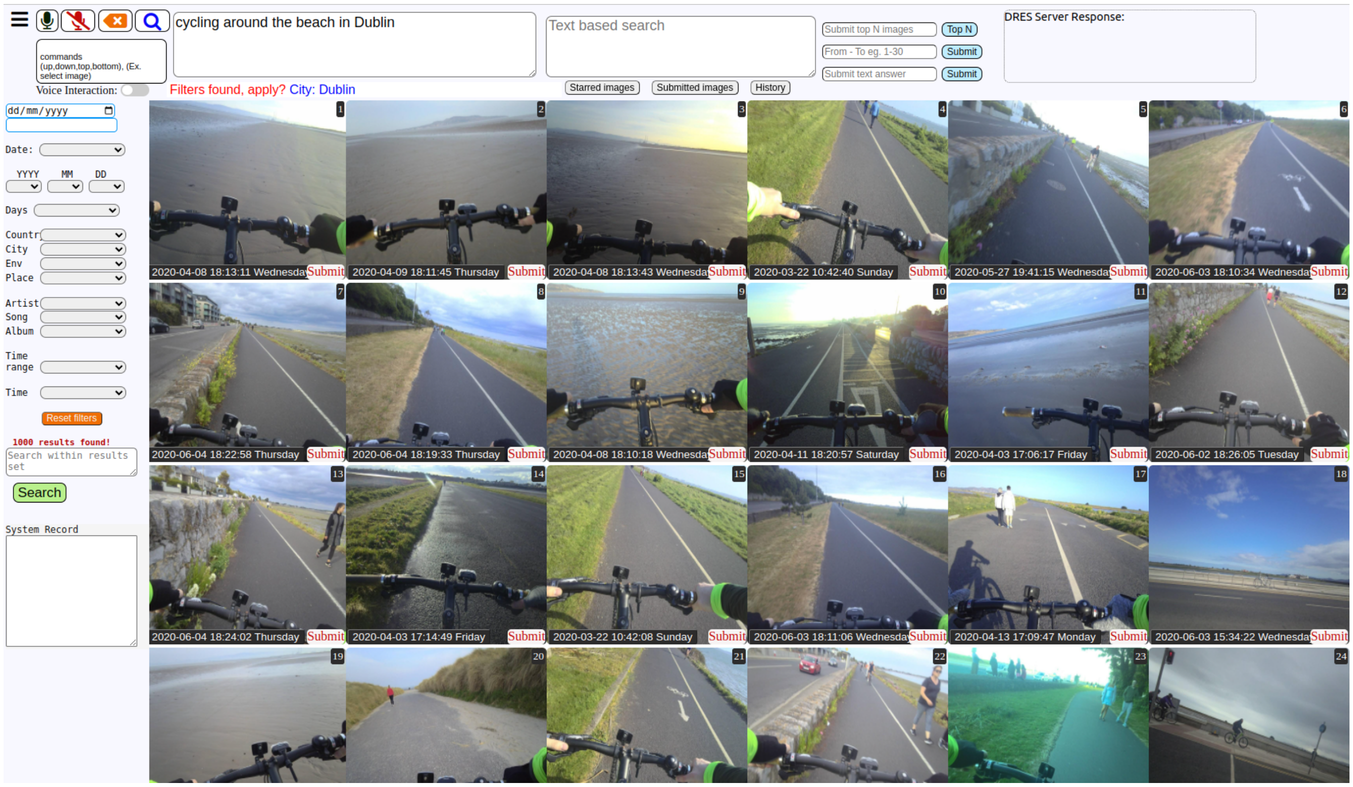

Figure 2: Voxento 4.0 Main Interface

continue to review other Speech Recognition and will validate them to determine if they are more suited to the current system. Recently, OpenAI has announced an Automatic Speech Recognition called Whisper API [17], which appears promising based on the reported features. Its evaluation is part of our current research where the goal is to configure their services to our system. Other configurations regarding the voice interaction were replicated from our previous versions [2], [3], [4].

In regard to the enhancements made to the voice interaction functionality this year, a logging feature was integrated which recorded and logged all voice interactions for analysis at a later date to enhance the features and performance of the next generation of the Voxento app. Additionally, new voice commands have been added to facilitate the new search topics, such as selecting a number of images for later submission to the evaluation server. In addition, the voice interaction was limited to certain aspects of the interface. We have improved the voice interaction with additional voice commands to cover these aspects.

5.2 Additional User Interface Features

The system interface has expanded significantly through the development of additional filters and elements. The redesign introduces improvements in the structure, content location, visual interface and style of image display. A general interface redesign was implemented based on experiences and feedback from users of the previous system. We believe users will now experience a better interaction and reach the important buttons and filters while viewing the images as the nature of the competition requires. Figure 2 shows the main interface. The setting of the voice interaction is inside a hidden panel until it is accessed by the user. It is assumed that this panel will not be accessed often. The top part of the interface is reserved for the query bar, voice interaction control, displayed message information and message centre. It also includes a specific text bar for submitting text answers using text based on the new changes in the LSC's search tasks. The left side of the interface integrates various search filters in a horizontal panel. In addition, we added a new option to select a range of images for submitting to the evaluation server, for topics that required many images to be submitted. We also implemented a feature to highlight (star) images of interest, by recording a history of starred images and submitted images. The interface has a text bar for searching within the results set returned from the server. We also expanded the search to cover most columns in the results set, from a previous version that was limited to semantic name information.

Additionally, interaction logging is implemented to capture all user interactions for later analysis. This involves voice and interface interaction and also the system log. A new feature is a scroll text bar containing live recorded speech recognition. Thus, it will help users to refer to the recorded speech and also for future analysis such as comparing the LSC queries with the users' speech query.

In order to keep the user interface intuitive for novices, we try to maintain a clear interface with few components that require complex interactions. We insist on this limitation as the eventual goal of Voxento is to provide a fully voice-controlled lifelog search tool. Other significant and valuable features contained in our previous system, such as dynamic results filters, calendar filters and detected filters from the query, new filters for the new LSC dataset are retained in our current version.

5.3 Search Engine Developments

The current back-end contains an additional text-based search engine in addition to a text-image embedding engine. We have enhanced the text-based search methodology to operate over all the LSC metadata. The Contrastive Language-Image Pre-training (CLIP) model [16] was also employed to support natural language queries as the main method for cross-modal visual retrieval. The maximum number of results returned for a query is set to depend on the search type, whether image-based search or text-based search. The backend server configuration is built in Python including the retrieval API, search functions, indexing and ranking. For detailed information on the development of the backend, please refer to [3] and [4].

In Figure 1, the final backend API will have two main search engine tasks, supporting two different query types in the main interface. Firstly, a text query based on visual image content will be converted to a vector representation using the CLIP model, which will then be compared to the image representation generated earlier using cosine similarity. As a result, high similarity values are sorted in a ranked list. Using the second query type, users can search in the prepared metadata, but limiting the search to specific aspects of the metadata, such as semantics or place names. The system will then display a list of top-matching images based on these semantic metadata results. These two search engines were already developed in the previous versions, and we have improved the retrieval functionality according to different search tests scenario done.

In the previous participation at LSC'22, we used the CLIP model version (ViT-L/14@336px), which was a significantly larger model than previous models. As a consequence, the size of the image embeddings becomes much larger because the volume of the LSC'22/23 dataset increased nearly 4 times compared to the previous dataset. As a result, retrieval functionality performance was adversely affected for Voxento at LSC'22. Based on feedback received from the Memento system [1], we used Facebook AI Similarity Search (FAISS) [13], which is a library to support fast searching through multimedia content. It supports different indexing methods and works by clustering the dataset index. As a result, the system's performance improved and can be estimated as 40% better but this is not enough for an efficient LSC system in the competition. We also explored Elasticsearch to find another technique to reduce the time of the retrieval functionality. However, as speed is a critical factor in assessing the system efficiently in the competition, we decided to use the lower CLIP model version which is (ViT-L/14). Based on examining the results of some test queries, we do not consider that this will have a major impact on our performance this year.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we presented an improved version of Voxento (4.0), our prototype voice-controlled interactive lifelog retrieval system. Voxento version 2.0 and 3.0 employing the CLIP model proved to be competitive to previous LSC'21 and LSC'22 competitions. Here, we presented a summarised improvement on the current version which will provide a better user experience targeting novice users at LSC'23. The improvement includes the dataset representation, main interface, search engine and voice interaction. For future work, this system is planned to be fully voice-based interaction. As illustrated in the paper we explored other speech recognition to be tested and deployed to the system.

ACKNOWLEDGMENTS

Research funded by Science Foundation Ireland under grant number SFI/12/RC/2289_P2 and also by the Ministry of Education in Saudi Arabia.

REFERENCES

[1] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2022. Memento 2.0: An Improved Lifelog Search Engine for LSC'22. In *Proceedings of the 5th Annual Lifelog Search Challenge (LSC '22)*. Association for Computing Machinery, New York, NY, USA, 2–7. https://doi.org/10.1145/3512729.3533006

[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). Association for Computing Machinery, 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). Association for Computing Machinery, 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. Association for Computing Machinery, New York, NY, USA, 43–47. https://doi.org/10.1145/3463948.3469071

[5] Aaron Duane and Bjorn THORNór Jónsson. 2021. ViRMA: Virtual Reality Multimedia Analytics at LSC 2021. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan). Association for Computing Machinery, New York, NY, USA, 29–34. https://doi.org/10.1145/3463948.3469067

[6] Jim Gemmell, Gordon Bell, and Roger Lueder. 2006. MyLifeBits: A personal database for everything. *Commun. ACM* 49, 1 (2006), 88–95. https://doi.org/10.1145/1107458.1107460

[7] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Duc Tien Dang Nguyen, Michael Riegler, Luca Piras, Minh-Triet Tran, Jakub Lokoč, and Wolfgang Hürst. 2019. [Invited papers] Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC2018). *ITE Transactions on Media Technology and Applications* 7 (04 2019), 46–59. https://doi.org/10.3169/mta.7.46

[8] 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

[9] Cathal Gurrin, Liting Zhou, Graham Healy, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoć, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Klaus Schöffmann. 2022. Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *Proceedings of the 2022 International Conference on Multimedia Retrieval* (Newark, NJ, USA) *(ICMR '22)*. Association for Computing Machinery, New York, NY, USA, 685–687. https://doi.org/10.1145/3512527.3531439

[10] 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)*. Association for Computing Machinery, New York, NY, USA.

[11] Silvan Heller, Luca Rossetto, Loris Sauter, and Heiko Schuldt. 2022. vitrivr at the Lifelog Search Challenge 2022. In *Proceedings of the 5th Annual Lifelog Search Challenge (LSC '22)*. Association for Computing Machinery, New York, NY, USA, 27–31. https://doi.org/10.1145/3512729.3533003

[12] Nhat Hoang-Xuan, Hoang-Phuc Trang-Trung, E-Ro Nguyen, Thanh-Cong Le, Mai-Khiem Tran, Tu-Khiem Le, Van-Tu Ninh, Cathal Gurrin, and Minh-Triet Tran.

2022. Flexible Interactive Retrieval SysTem 3.0 for Visual Lifelog Exploration at LSC 2022. In *Proceedings of the 5th Annual Lifelog Search Challenge (LSC '22)*. Association for Computing Machinery, New York, NY, USA, 20–26. https://doi.org/10.1145/3512729.3533013

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

[14] Emil Knudsen, Thomas Holstein Qvortrup, Omar Shahbaz Khan, and Björn THORNór Jónsson. 2021. XQC at the Lifelog Search Challenge 2021: Interactive Learning on a Mobile Device. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan). Association for Computing Machinery, New York, NY, USA, 89–93. https://doi.org/10.1145/3463948.3469063

[15] Andreas Leibetseder, Daniela Stefanics, and Klaus Schoeffmann. 2022. lifeXplore at the Lifelog Search Challenge 2022. In *Proceedings of the 5th Annual Lifelog Search Challenge (LSC '22)*. Association for Computing Machinery, New York, NY, USA, 48–52. https://doi.org/10.1145/3512729.3533005

[16] 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

[17] 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

[18] 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. Association for Computing Machinery, New York, NY, USA, 38–42. https://doi.org/10.1145/3512729.3533008

[19] 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)*. Association for Computing Machinery, New York, NY, USA, 32–37. https://doi.org/10.1145/3512729.3533012