Voxento 3.0: A Prototype Voice-Controlled Interactive Search Engine 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

Voxento is an interactive voice-based retrieval system for lifelogs which has been redeveloped and optimised to participate in the fifth Lifelog Search Challenge LSC'22, at ACM ICMR'22. Based on the previous experience in the LSC competition and ranked in the top 4 in the last LSC'21 competition among 17 participants, we present a revised version of Voxento to address the critical points to improve the efficiency of retrieval tasks in lifelog datasets. Basically, Voxento provides a spoken interface to the lifelog data, which facilitates an expert and novice user to interact with a personal lifelog using a range of vocal commands and interactions. Briefly, we made some important improvements to support both the retrieval of content and system interaction. This latest version has been enhanced with the addition of a text-based search feature, new filters based on new metadata provided in lifelog data, rich visual information and features and enhanced speech query. Also, the data preparation tasks comprised a new function to reduce the number of non-relevant images and the latest CLIP model version used to derive features from images. The long term development of Voxento includes a lifelog retrieval that supports speech and conversation interaction with less physical actions required by users such as using a mouse. The system presented here uses a desktop computer in order to participate in the LSC'22 competition with the option to use voice interaction or standard text-based retrieval.

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. 2022. Voxento 3.0: A Prototype Voice-Controlled Interactive Search Engine for Lifelogs. In *Proceedings of the 5th Annual Lifelog Search Challenge (LSC '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3512729.3533009

LSC '22, June 27–30, 2022, Newark, NJ, USA

ACM ISBN 978-1-4503-9239-6/22/06.
https://doi.org/10.1145/3512729.3533009

1 INTRODUCTION

In simple terms, lifelogging is the action of capturing most aspects of an individual's life experiences by using a group of specific technologies and software to gather data such as wearable cameras, fitness trackers, sleep trackers, smartphones, smartwatches and various other applications and products such as Google Glass [6]. There are factors behind the increasing popularity of lifelogging such as the development of efficient and low-cost storage and the availability of sensor devices [6]. Recently, many people have used phones to record and share their life's experiences and moments with others through various social applications. This can also be considered as a form of lifelogging. Lifelog data, generated from wearable cameras and other technologies requires re-organization, re-structuring and integration into a specific form, to support efficient information retrieval. This is due to the heterogeneity of sensors and technologies used in generating this data. For instance, MyLifeBits [5], the first lifelog retrieval system prototype, had different types of data such as emails, diary, images, audio "recording voices and music", video, TV and radio channels and documents. In lifelog data, images represent the majority of objects with associated metadata. Also, lifelog data usually contains biometric data such as physical activities, heartbeats, location data including country and semantic names, dates and times.

The Lifelog Search Challenge (LSC) is an annual benchmark competition workshop to tackle challenges in lifelog retrieval tasks [7]. LSC invites researchers to develop an interactive lifelog retrieval system and participate in a live competition. At LSC'21, the workshop was publicly available to watch by streaming platform and public audiences were invited to watch the competition. From the start of the LSC workshop in 2018 until 2021, the number of participants has increased to 17 participants. LSC workshop has been running for 4 years successfully, and it is now in its fifth year. The organizers of LSC have set up a workshop called *LSC Planning Workshop* to discuss the progress of the LSC workshop. It occurred in November 2021 in physical format. A result of this workshop was a new and larger dataset, as well as the integration of the LSC workshop deeper into the ACM ICMR conference programme (Grand Challenge 2022). In terms of the LSC workshop process, the details of how the workshop runs, how to evaluate the participant's system performance and the different query tasks are described in [7].

In this paper, we present the third version of Voxento, a prototype voice-controlled interactive retrieval system for lifelogs. Voxento has been redeveloped and enhanced in order to reach a stage where users/lifeloggers can communicate by voice as a conversation with a personal lifelog retrieval system to retrieve specific moments and

find relevant images or information. The goal of our research is to enhance the interaction and accessibility of lifelog data using voice as a form of communication, which is a concept similar to Google home or Amazon Alexa. Voice-controlled lifelog retrieval still has not gained much attention and this motivates us to work on developing and enhancing the system. Voxento participated for the first time in LSC'20 [2], focusing on introducing voice interaction and had a standard MongoDB database. In LSC'21, Voxento version 2 [3] included major enhancements in voice interaction, filtering and a newly developed retrieval API provided by [1]. The system proved to be competitive and was ranked the 4th top-performing system. Voxento 2.0 solved 18 out of 23 search tasks with only 3 wrong submissions, which represented the lowest compared to other teams.

In the system presented here, we have enhanced the speech query detection, added text-based search for finding relevant images based on semantic names, added filters for new metadata such as artist and song fields. We also rebuilt the backend retrieval API. Regarding data processing, we use the CLIP model [13] to derive features from each image. The list of image features is subsequently used to find the similarity between text and images and ranks those images which best represent the appropriate results. This year's lifelog dataset is the largest since the start of the LSC workshops. Thus, one key contribution of this research is to provide solutions to reduce the size of the lifelog data through better analysis and processing of the LSC dataset. Finally, all existing features in previous versions are retained in this latest version.

2 RELATED WORK

Since the Lifelog Search Challenge first ran in 2018, there have been many different approaches and systems developed during that time. In addition, various systems have participated more than once such as lifeXplore [10], Exquisitor [8] and Myscéal [16]. Some systems have been rebuilt with different concepts but share similar platforms, such as Virtual Reality in vitrivr-VR [15] and ViRMA [4] inspired by the M^3 data model. An example of a new system developed with a new idea is XQC [9], an interactive learning on mobile phones that supports IOS, Android and web platforms. This system uses Exquisitor's backend server [8]. Surprisingly, although a small screen can be a challenge in the LSC competition, XQC [9] with its features performs slightly better than the Exquisitor system and in some aspects, demonstrates a similar performance.

In the last LSC competition, there were many systems that performed well and here, we highlight the top performing retrieval systems at the LSC'21 competition. The top four performing systems were close in terms of the overall score. The system which performed best was Myscéal [16], which also won the first in the LSC'20 competition. Myscéal has been revised and upgraded with the improved features inspired by novice users experiments. The second-best system was SOMHunter [11], which also won at LSC'20. SOMHunter was an adaptation of a video search engine. In this extended version, it integrated the new CLIP text search model. The third ranked system was LifeSeeker [12] which participated three times in 2019, 2020 and 2021. It focused on searching and filtering by text query using a weighted Bag-of-Words model with visual concept augmentation and weighted vocabularies.

Another new system which was ranked 6th is called Memento [1]. This system was developed to address the challenge of interactive lifelog retrieval based on two aspects: firstly, bridging the semantic gap between queries and images; and secondly, supporting the efficient searching/browsing of the lifelog data. It is worth noting that our Voxento system and Memento share the same backend with some minor modification to support Voxento interaction and filters features. The developers of the Memento system paid special attention to preparing and processing the lifelog dataset, deriving features, making event segmentation, enhancing metadata and using CLIP model. We rebuilt the backend by considering these enhancements as well further improvements in interaction and visual interface and creating text-search based features. We observed that two LSC'21 systems, ViRMA [4] and PhotoCube [14] used the M^3 model, which defines the dimensions of a hypercube. A M^3 model refers to a multidimensional space that organizes media items into related groups to allows the user to explore media collections of relevant images. All participating systems placed an emphasis on supporting easy querying by users, enhancing the metadata to support information retrieval, with many systems implementing different filter types and a few used the CLIP model to generate images features.

3 VOXENTO'S PERFORMANCE AT LSC 2021

Voxento 2.0 at LSC'21 [3] performed well and was a competitive system in contrast to the previous participation at LCS'20. The system was optimised to use a desktop with mouse and keyboard at the competition. However, participation at both LSC'20 and LSC'21 provided valuable input to enhance the search engine. Based on experiences at LSC and some testing, we can now articulate a number of enhancements for this LSC event that we believe delivers a significantly superior system.

- **Search Filters**: the implemented filters were efficient in retrieving and narrowing the ranked results. However, the time filter was only filtering by hours and not as range. Thus, every time, we need to reset the results' panel to choose another hour. We redeveloped *time* as a range filter to be more specific when searching using time. Another important filter is semantic name. There were some queries which contained the semantic name and filtering using this feature locates the appropriate images more easily.
- **Search Engine**: in the previous version of Voxento [3], the integrated backend used the CLIP model for matching based on similarity. The ranked list contained appropriate images and results are promising. However, there were some queries for which we could not initially locate the relevant images and it was necessary for the user to browse and search for that specific moments manually, and sometimes it was necessary to reformulate the query. To support this task, we implemented a text-based search and with the support of the metadata, we could find the relevant images. This will have significant improvements for semantic name, artist name or song words.
- **Visual Interface**: here, we mean a specific aspect of the visual component, which is the image display. The image only displayed the number label representing the ranked

order and when the image was clicked, it then displayed a window showing more metadata with buttons including the submit button. Hence, during the search process at the LSC'21 search tasks, we sometimes found the relevant image but needed to click on the image in order to submit it. These were examples where we were uncertain and needed to click on the image to read the metadata. To solve and enhance the submitting of a relevant image, we now include a submit button in each displayed image as a label, together with basic image metadata such as date, time and semantic name.

- **User Interaction**: in future research, we are planning to do novice experiments to identify any enhancements required in terms of interaction, but have already made a minor improvement in the speech detection for the query. Based on the fast retrieval of results, we made the decision to submit the query to the server while the user was speaking the query. For example, when the user says the query, *"find a moment when I was taking a picture by my phone at a lake"*, the system will submit the query while the user speaks and will update the results whenever a query has changed.

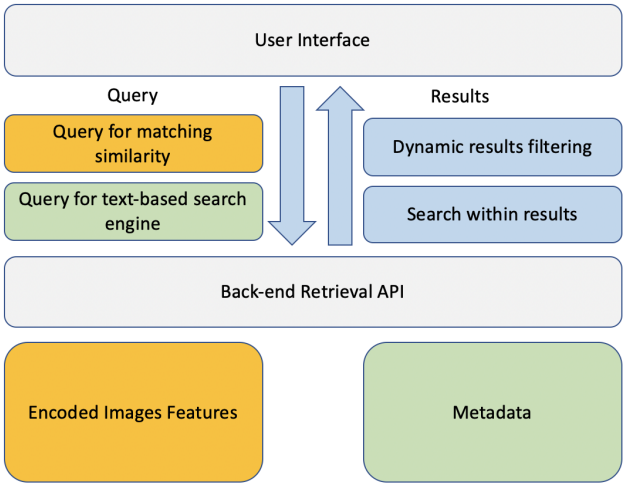

Figure 1: Overview System Architecture

4 LSC'22 DATASET

The lifelog dataset provided for this year's event is far larger than previous LSC workshops. We now briefly highlight some new aspects of the LSC'22 dataset in this section, with a more detailed description available in [7]. The LSC'22 dataset provided for the 5th Lifelog Search Challenge represents an 18 month period during the years 2019 and 2020, captured by one active lifelogger using wearable cameras. The lifelog dataset contains about 725k images. LSC organizers provided one month's data to participants at the outset with the final dataset made available about two months before the workshop. The full dataset included anonymised 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. New information provided for LSC'22 is image tags and text OCR captions.

As we rebuilt the backend API, motivated by [1], we also attempted to organise and process the data similar to the previous structure in order to match the frontend configuration and structure. We also made a contribution to enhance the metadata by reducing the number of images. Below are the summary tasks of processing data:

- **Data Enhancement**: We do not exclude the new caption information and instead use them for text-based search. In addition, we enrich the data by deriving features from existing metadata such as time (hour:minute), Date(day, month, year) and day name (Sunday, Monday, ... etc). This information can be extracted from the minute_id or UTC_time attributes. Moreover, in our previous system, the dataset already included blurred image label and after some testing, we found that some blurred images were relevant images for some queries. Thus, we decided to use the OpenCV library to identify a specific degree of blurring where we can see that these images no longer have values because they have a high degree of image blur. Another task of enhancing the metadata is to find images that have no value, to exclude them. We designed a function to identify images that contain high rates of black or white colour compared to the image size because sometimes the wearable camera is left on the table without turning it off. This means that it takes pictures of a roof or light and sometime left in a pocket or another dark location. The final task was to identify the duplicated images to exclude the worst and use only the best one. When exploring the one-month dataset, we found about 20 duplicated images that occupy space without any additional benefit.
- **Extracting Images Features**: Using the CLIP model [13] added a high value to the performance of our previous system. Thus, in this version, we retained the CLIP model to encode images into high-dimensional representations to detect rich scene semantics. By using the model, we compare the encoded query sentence with the encoded image features to find the matching similarity and then sorted them based on high similarity.
- **Event Segmentation**: Event segmentation was a helpful process during the LSC competition to follow queries containing different activities such as check in hotel and then drive a car. This methodology, used in a previous system [3], has now been integrated into our system. Segmentation of each day was based on events by identifying the current activity and the location name when the lifelogger performs some movement. Event segmentation can inform temporal sequencing of result documents using the interface.

5 OVERVIEW OF VOXENTO 3.0

In this section, we present an overview of the revised Voxento system and architecture with a detailed description of the main components. Voxento 3.0 is based on the previous version with a new text-based search engine to support the text-image search

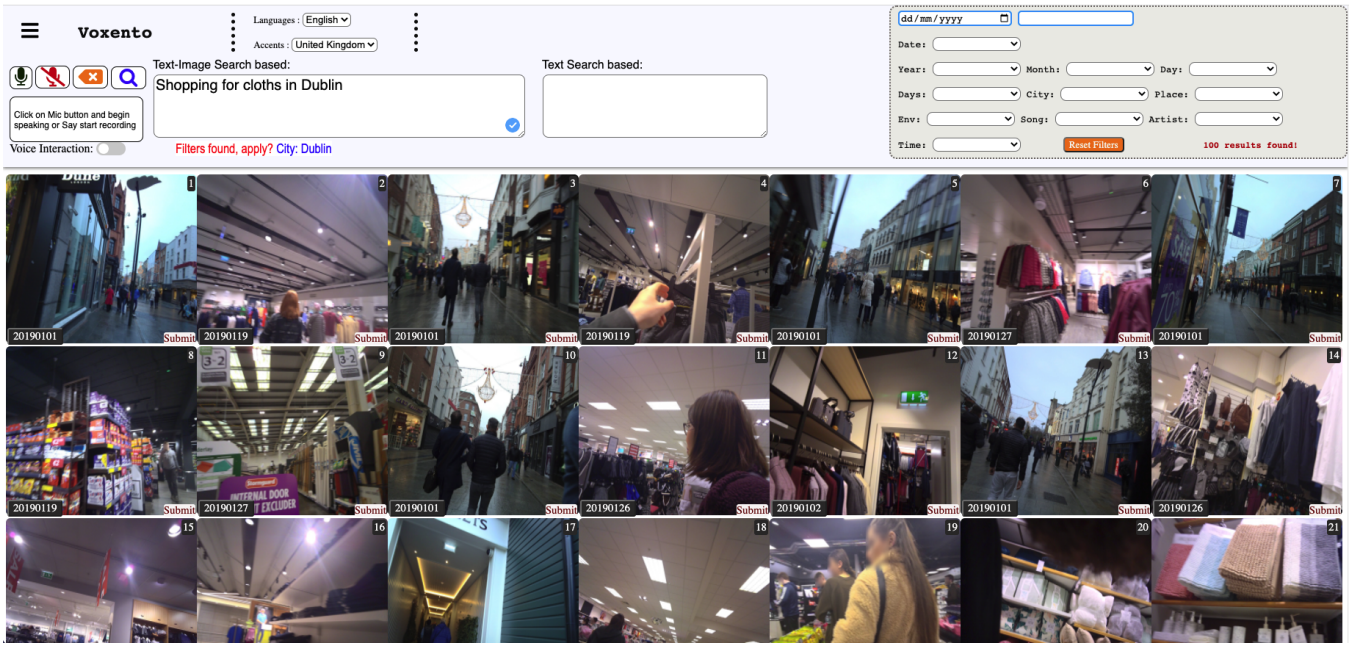

Figure 2: Voxento Main Interface

engine. In addition, the enhancements to the visual interface and filters are discussed in section 5.2. The architecture can be viewed as having four main components as shown in Figure 1 and 3: the user's voice and interface interaction, visual interface, and the backend API that has two search engines. Voxento version 3.0 has a similar interaction implemented but enhanced over previous versions.

5.1 Voice-based Implementation

For this version, the voice interaction functionality has no major changes and for details as to how the voice interaction receives commands and performs action, refer to [3]. Briefly, the Voxento interface was developed as a web-based application and the user uses a headset for voice interaction or mouse and keyboard for a standard retrieval system. For basic interaction, on system loading, the voice interaction feature automatically engages and awaits the user's command. The user can either utter a spoken command or revert to using a mouse and keyboard to interact with the interface. The user can swap between these two interaction methodologies. The Google Web Speech API is used through Chrome browser to support voice interaction. This enables speech recognition and synthesis through the web browser, with the language set to English with a standard accent. Other configurations were replicated from our first version [2].

The new feature in query speech detection is that when the user start speaking, the query and transcription are shown as live. Thus, the results automatically submit each time the query changes. For instance, when the user starts a query such as *"find a moment when I was driving the car*, the system will submit this query and when the user adds more words in the query or continues the query, the system submits every time, the updated query rather than having to stop the recording then submit the final or complete query.

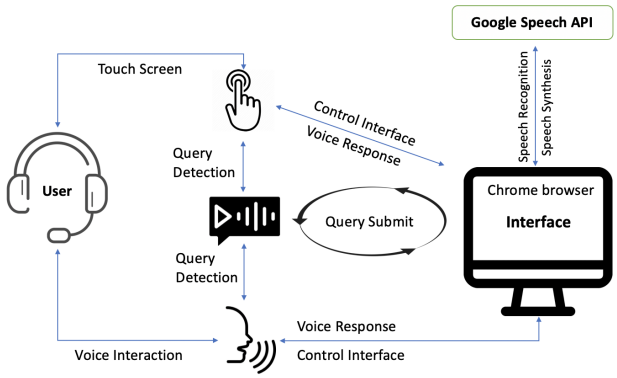

Figure 3: An Overview of System Interaction

5.2 User Interface

Since the development of the first version, we separated and positioned all control functions to remain inside the top "query panel" and thus, leave space for the visual interface. We try to always maintain a clear interface to keep it simple for novice users. In addition, the voice interaction has a clear presentation including different colours, for instance, *start* and *stop* recording. The design and the structure of the system is illustrated in Figure 2 with the interface explained in [3].

All of the features contained in our previous system, such as dynamic results filters, calendar filter and detected filters from query, are retained in our current version. However, we have implemented some new filters. The first filter is an *environment* filter, such as Work or Home, based on recognizing the need to use this filter. The

second filer uses the semantic name such as a restaurant's name where the system can face difficulty in locating the relevant images. Note that it can be missed by using CLIP unless the semantic name is represented in the image. We fixed the issue of filtering time using the drop-down option *hours* and the range feature to narrow the scope. We also created filters based on the updated metadata to include searches using, for example, artists and songs words.

5.3 Search Engine

The backend was rebuilt to include a text-based search engine in addition to a text-image search engine and with minor modifications to call requests to support Voxento's interaction requirements. The CLIP model [13] was used to support natural language queries. A challenge for the CLIP model is that when a query has, for instance, a city name like Dublin or London, the results will not be accurate. Therefore, the system already implemented a feature to exclude these words and use them as filter words from the query. The backend API was built using the Flask python framework using RESTful API functionality. The maximum number, N, of results returned for a query, is set to 1,000. The backend API was evaluated in LSC'21 and showed that the system was capable of detecting 18 out of 23 LSC'21 topics, so we expect it to be very competitive with these new features at LSC'22. The evaluation results showed that our system achieved the highest precision among all participants.

The backend API has two main search engine tasks, supporting two different text queries in the main interface. The first text query is for a search based on images so the backend will take the query and convert it to a vector representation using the CLIP model which then will be compared to the image representation, generated previously, using cosine similarity. This results in the ranked list sorted by high similarity values. The server then uses the images list to fetch the image metadata, which is also located in the server. The second text query is used when there is a need to search for a specific text in the database, such as semantic or place name. Finally, for both text queries, the results are sent as a JSON response to the user interface.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we presented an enhanced version of Voxento, our prototype voice-controlled interactive lifelog retrieval system. Voxento 2.0 proved to be competitive and performed well at the previous LSC'21 competition. We presented a summarized experience at LSC'21 and articulated the improvements needed for a better system, which were mainly in dataset preparation, search filers and engines and interaction. We presented a system overview and interface, with a detailed description of new enhancements. Regarding future work, we believe that Voxento can be further developed to be a standalone system or application that can enhance the accessibility. Our current plan is to move to a fully wearable platform or migrate from the desktop environment to become a truly pervasive computing environment.

ACKNOWLEDGMENTS

We acknowledge the support of Science Foundation Ireland and the Insight Centre for Data Analytics through the grant number SFI/12/RC/2289-P2 and the Ministry of Education in Saudi Arabia for sponsoring the PhD research of the primary author.

REFERENCES

[1] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2021. Memento: A Prototype Lifelog Search Engine for LSC'21. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan). Association for Computing Machinery, New York, NY, USA, 53–58. https://doi.org/10.1145/3463948.3469069

[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] 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

[5] 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

[6] 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

[7] Cathal Gurrin, Liting Zhou, Graham Healy, Björn Por Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Klaus Schöffmann. 2022. Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *Proc. International Conference on Multimedia Retrieval (ICMR'22).* Association for Computing Machinery, New York, NY, USA.

[8] Omar Shahbaz Khan, Aaron Duane, Björn THORNór Jónsson, Jan Zahálka, Stevan Rudinac, and Marcel Worring. 2021. Exquisitor at the Lifelog Search Challenge 2021: Relationships between Semantic Classifiers. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan). Association for Computing Machinery, New York, NY, USA, 3–6. https://doi.org/10.1145/3463948.3469255

[9] 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

[10] Andreas Leibetseder and Klaus Schoeffmann. 2021. LifeXplore at the Lifelog Search Challenge 2021. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan). Association for Computing Machinery, New York, NY, USA, 23–28. https://doi.org/10.1145/3463948.3469060

[11] Jakub Lokoč, František Mejzlik, Patrik Veselý, and Tomáš Souček. 2021. Enhanced SOMHunter for Known-item Search in Lifelog Data. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan). Association for Computing Machinery, New York, NY, USA, 71–73. https://doi.org/10.1145/3463948.3469074

[12] Thao Nhu Nguyen, Tu Khiem Le, Van Tu Ninh, Minh Triet Tran, Nguyen Thanh Binh, Graham Healy, Annalina Caputo, and Cathal Gurrin. 2021. LifeSeeker 3.0: An Interactive Lifelog Search Engine for LSC'21. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan). Association for Computing Machinery, New York, NY, USA, 41–46. https://doi.org/10.1145/3463948.3469065

[13] 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

[14] Jihye Shin, Alexandra Waldau, Aaron Duane, and Björn THORNór Jónsson. 2021. PhotoCube at the Lifelog Search Challenge 2021. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan). Association for Computing Machinery, New York, NY, USA, 59–63. https://doi.org/10.1145/3463948.3469073

[15] 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). Association for Computing Machinery, New York, NY, USA, 17–22. https://doi.org/10.1145/3463948.3469061

[16] Ly Duyen Tran, Manh Duy Nguyen, Nguyen Thanh Binh, Hyowon Lee, and Cathal Gurrin. 2021. Myscéal 2.0: A Revised Experimental Interactive Lifelog Retrieval System for LSC'21. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan). Association for Computing Machinery, New York, NY, USA, 11–16. https://doi.org/10.1145/3463948.3469064