LifeSeeker 6.0: Leveraging the linguistic aspect of the lifelog system in LSC'24

Hoang-Bao Le
Thao-Nhu Nguyen
Tu-Khiem Le
School of Computing, Dublin City University
Ireland

Minh-Triet Tran
Thanh-Binh Nguyen
Ho Chi Minh University of Science - Vietnam National University
Vietnam

Van-Tu Ninh
Liting Zhou
Cathal Gurrin
ADAPT Centre, School of Computing, Dublin City University
Ireland

ABSTRACT

Supporting effective access to digital lifelogs is a challenging research task because of both the volume and variety of multimodal lifelog data, as well as the many and diverse types of information need that should be supported. In this paper, we introduce a new version of LifeSeeker called LifeSeeker 6.0, for the 2024 edition of the ACM Lifelog Search Challenge. Our enhancements include the improvements to the user interface and the backend reconstruction by combining the E-LifeSeeker structure with using contrastive learning between texts. These adjustments are aimed at accelerating the correlation between the huge image collection and the text input, thereby enhancing the retrieval accuracy and efficiency.

CCS CONCEPTS

• **Information systems** → **Multimedia databases**; • **Users and interactive retrieval**; • **Search interfaces**; • **Human-centered computing** → **Interactive systems and tools.**;

KEYWORDS

lifelog, information retrieval, interactive system, clip, contrastive learning.

ACM Reference Format:
Hoang-Bao Le, Thao-Nhu Nguyen, Tu-Khiem Le, Minh-Triet Tran, Thanh-Binh Nguyen, and Van-Tu Ninh, Liting Zhou, Cathal Gurrin. 2024. LifeSeeker 6.0: Leveraging the linguistic aspect of the lifelog system in LSC'24. In *The 7th Annual ACM Lifelog Search Challenge (LSC '24), June 10, 2024, Phuket, Thailand*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3643489.3661121

1 INTRODUCTION

With the advent of smartphones and similar devices, lifelogging has become much more accessible. Lifelogging is the passively capturing and data-storing process what we see during the whole day. Thanks to the devices' support, we have a great opportunity to not only remind us of visual memories but also track important information from the past. In order to prompt the lifelogging retrieval and take advantage of the recorded images, there are some organised activities such as NTCIR-Lifelog[1], ImageCLEF Lifelog[2] and the challenge for this paper - Lifelog Search Challenge[3] (LSC). Besides that, our lifelog data can be stored as video format, therefore, an annually special competition named as Video Browser Showdown[4]. The primary function of these activities is prompting the system development based on their accuracy, speed, easy interaction, and multitasking adaptability.

In the latest LSC'23[4], there are three main tasks: Known-item search, Ad-hoc, and Question-Answering (QA). While the known-item search task forms the fundamental task in lifelog retrieval for a long period of time, the remaining two tasks have been introduced since 2022. In the Ad-hoc task, competitors aim to submit as many correct images as possible with the related input query. For example, with the request '*Find examples of when I was eating avocado for breakfast.*', the system is targeted to return all the images containing the avocado dish in the morning. On the other hand, the QA task requires the exact answer to the corresponding question. For instance, in a QA task, the system compulsorily understands the contextualised process to answer questions such as '*Which airline did I fly with most often in 2019?*' or '*For a while in 2019, I wrote a number on my office window in red pen. What was that highest number that I wrote?*'.

First introduced in 2019, LifeSeeker has been released in five versions [13, 14, 17–19]. Through the years, our target is utilising state-of-the-art vision language models and improving the user interface for LifeSeeker.

In this paper, we release the latest version - LifeSeeker 6.0for LSC'24 [9], which mainly concentrates on the linguistic aspect by applying the contrastive learning. Instead of only matching the image data and the text query, we supplied the detailed description combined with the metadata (time and location) and identified the correlation not only between image-and-text (IaT) but also text-and-text (TaT). Additionally, we redesigned the interface by supplying the description and the possible answer if the input is the question for every image.

2 RELATED WORKS

Since first organised in 2018, a large number of systems have been released and prompted the development of LSC solutions. Last year, LSC attracted 11 teams participating in and the top three

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

ACM ISBN 979-8-4007-0550-2/24/06
https://doi.org/10.1145/3643489.3661121

[1] http://lifelogsearch.org/ntcir-lifelog/NTCIR17/
[2] https://www.imageclef.org/2020/lifelog
[3] http://lifelogsearch.org/lsc/
[4] https://videobrowsershowdown.org

ranked systems are lifeXplore [28], MyEachtra [31] and Memento 3.0 [1]. Firstly, lifeXplore, which has been released first time since 2018, has the backend combination of Object Detection (YOLOv7 [33]), Concept Detection (EfficientNet [30]), Text Detection (CRAFT [3]) and Video-Text Embedding (OpenCLIP [6]). MyEachtra has a highlight enhancement with the "event-based" approach that prevents users from missing the relevant images and retrieves them quickly and accurately. Moreover, Memento 3.0 creates the clusters for the whole image embeddings and then matches the search query with the best cluster centroids. By implementing the cluster-based search technique, the system lowers considerably the query processing time by nearly 75%.

Besides the top three models and ours, there are other noticeable systems participating in the LSC'23. In addition to using CLIP [21], Voxento 4.0 [2] focuses more on improving the systems's interface and using Whisper API [22] for the voice interaction. With the second version of MEMORIA [24], Ribeiro et al. applied the free-text search that the pipeline in their system divides the raw text into objects, events, activities, locations, dates, times, and temporal aspects, and created a database query after using word2vec [16] for expanding the found tokens. For the first participation, instead of applying CLIP, MemoriEase [32] uses BLIP [15] for the embedding-based retrieval approach to reduce the semantic gap between images and text queries. Furthermore, Tran et al. [32] also combined BLIP with concept-based retrieval approaches, implemented through a search engine on Elastic Search. In the other way, vitrivr [29] is designed to focus on the user interface more and this system has three different types of view such as the cylindrical result view, the detail view and the multimedia drawer.

LifeLens [11] is designed to be minimalist and user-friendly, which priorities more on the image features and provides a more intuitive look as same as simple using. In addition, LifeGraph 3 [26] identified three clusters as temporal, spatial and visual so as to arrange sequences of the Lifelog entries into meaningful bins. By implementing the Large Language Models (LLMs), Lifelog Discovery Assistant or FIRST 3.0 [10] implemented the few-shot learning prompts to assist users the hints and support them a number of options for modifications. As same as the idea of MemoriEase, however, LifeInsight [20] is added two highlight points as the function of explicite relevance feedback to re-rank the results, and Roccio algorithm [25] to form the initial query vector or modify the previous input query vector for searching or re-ranking.

Overall, a majority of the systems attending in LSC'23 used the visual-language embeddings from CLIP and added particular techniques to enhance the search. Our proposed system, LifeSeeker 6.0, focuses on the language aspect by being added the detailed description for the query matching.

3 SYSTEM OVERVIEW

3.1 LSC Dataset

The LSC'24 dataset is as same as the previous LSC'23 and LSC'22 datasets. Being collected by a Narrative Clip[5] device, this multimodal dataset provides a comprehensive view of a single lifelogger's experiences for 18 months (2019-2020). The challenge dataset is comprised of three components:

- **Core image dataset**: The images, extracted from the wearable in 1024×768 resolution, have undergone thorough redaction and anonymisation. To not only protect the lifelogger's privacy but also ensure that other individuals appearing in the images remain anonymous, all the human faces, as well as the readable contents (phone and laptop screens, laptop displays, or documents), have been blurred or removed if possible.
- **Visual Concepts**: For every image, the visual concept data provides valuable insights into the object information (objects, scenes, and context) within the corresponding image and the text detected by using the OCR models.
- **Metadata**: The metadata contains the necessary information related to the user's time (date, local time, and part of the day) and location (GPS, coordinates, city, and country) at the moment the image is captured.

3.2 Model Structure

Following the last version of LifeSeeker [17], we adopt the two-stage structure including offline and online as illustrated in Figure 1. During the offline stage, we create an embedding vector space for every image and its corresponding caption. The generated caption is enriched with metadata consisting of spatial, temporal, and other necessary information. Next, the system embeds the users' input query and then matches it with the aforementioned space by using the cosine similarity score. The results are returned based on the final score:

$$\text{Top\_N\_matches} = W_{IaT}\text{sim}_{IaT} + W_{TaT}\text{sim}_{TaT}.$$

where sim_i is the cosine similarity score of item i and $W_{IaT}, W_{TaT} = 1$ denotes the weights for the Image-and-Text and Text-and-Text similarity, responsively. We propose that $W_{IaT} + W_{TaT} = 1$, and $W_{IaT} \geq W_{TaT}$ as we aim to highlight IaT similarity more than the TaT one. The higher the final score is, the more relevant the image-query pair is. In our system, we set the weight as $(W_{IaT}, W_{TaT}) = (0.7, 0.3)$.

Furthermore, we also perform the combined description (Cap-Meta - *Cap*tion *Meta*data) for the images on the user interface alongside the filter options for the time, location, objects, or text visible in the images (like part of the day, at work, etc.) to the current query.

3.3 Contrastive Learning Underlying Structure

Contrastive learning is a deep learning technique that teaches the model to learn the general features of dataset without labels by distinguishing between pairs of similar or dissimilar data points. The technique has driven a revolution in not only learning visual representations, powering methods such as SimCLR [5], CLIP [21], and DALL-E 2 [23], but also sentence similarity such as CERT [7] and SimCSE [8].

In the offline stage (figure 2, we utilise contrastive learning twice in order to enhance the correlation between the image with its information and the input query or the question.

- **Image-and-Text (IaT)**: Inheretting the success of the last two versions of LifeSeeer, LifeSeeker 6.0 is developed based on the same structure that is applied the visual language

[5]http://getnarrative.com

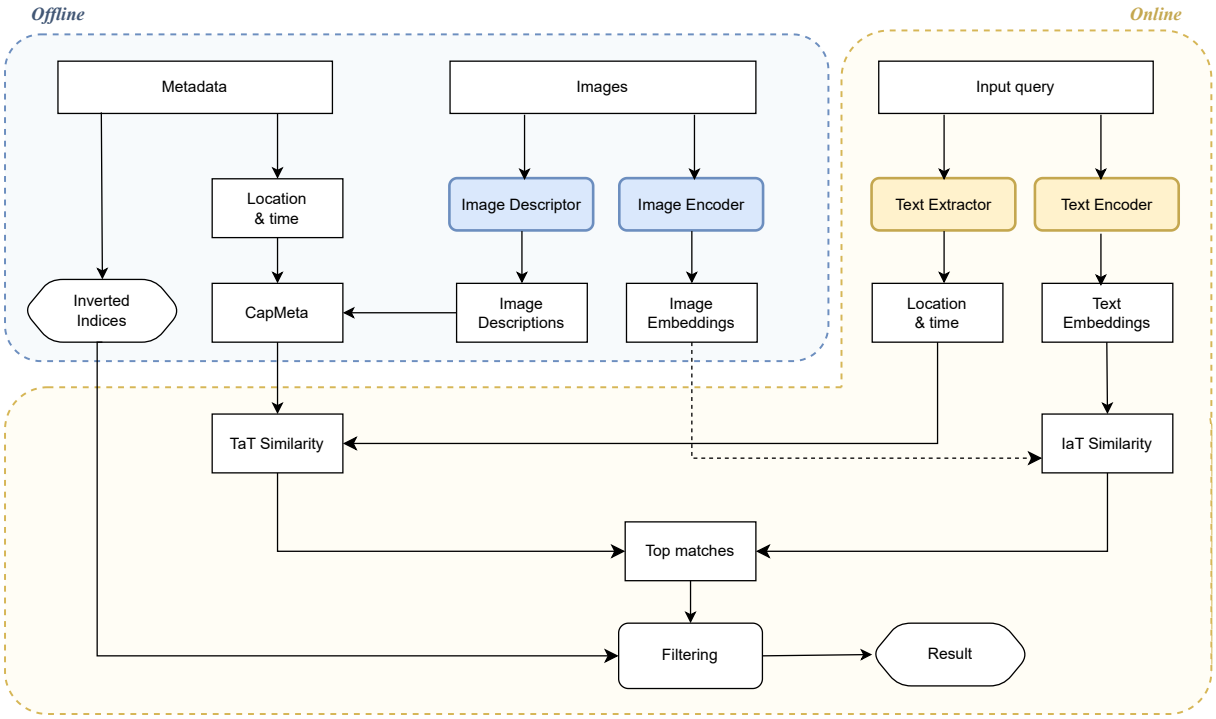

Figure 1: The System Architecture of LifeSeeker 6.0.

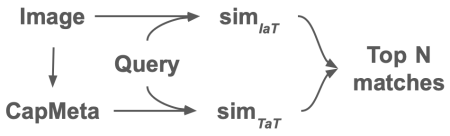

Figure 2: The underlying structure of Life6ker.

embedding models such as CLIP [21], CoCa [34], BLIP [15] and ALIGN [12].

- **Text-and-Text (TaT)**: To implement the contrastive learning between text and text, firstly, we use the model FUSECAP [27] to generate a first-sight description for each image. The model FUSECAP[6] contains three frozen visual experts including an object detector, an attribute recogniser, and an Optical Character Recogniser (OCR) (figure 3).

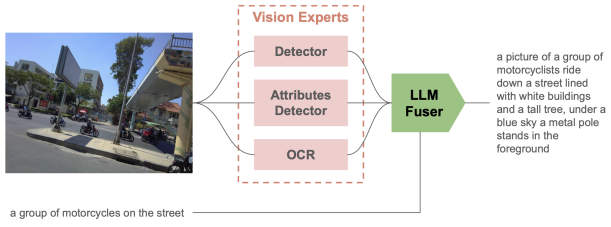

Figure 3: The idea inside FUSECAP.

Additionally, FUSECAP is pre-trained on 12M image-enriched caption pairs with a captioning generation BLIP-based model. Therefore, this model has the ability producing more precise and detailed descriptions. Moreover, we supply the spatial and temporal information as the following format

$$\text{CapMeta} = \textsc{concat}(\text{description}, \text{time}, \text{location})$$

to make it more detailed (example: figure 4). By doing this, it can enhance the image content and match the query better. As QA is a special task which requires the text answer, we also adjust to gain a better answer. Firstly, we train the model to detect keywords inside the question and the system returns the top related images. Besides the description of the images, we also set the system to recommend the possible answer based on the question by using the model[7] BLIP pre-trained on QA task. We believe that with these adjustments, the users will not only get a more intuitive interface but also have multiple options of answers for the given question.

3.4 User Interface Adjustments

As this year we concentrate more on developing the backend structure, we reuse the frontend design of E-LifeSeeker [17] leverage the current intuitive user interface with an addition of a small number of enhancements. Precisely, the LifeSeeker 6.0's user interface was equipped with four components: the free-text search and filter box, the automatic question generation display, the search progress

[6]https://huggingface.co/noamrot/FuseCap_Image_Captioning

[7]https://huggingface.co/Salesforce/blip-vqa-capfilt-large

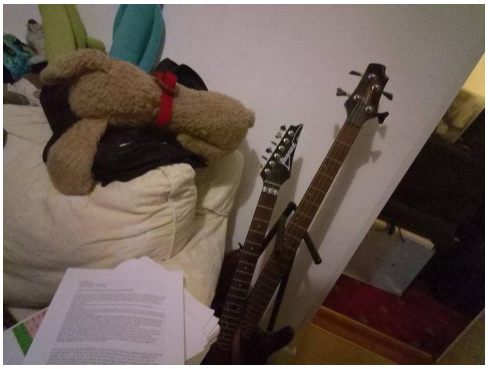

Figure 4: An example of CapMeta - *a picture of a teddy bear sits on a bed next to a brown guitar and an open book, with a white wall in the background # Time: 2020-01-01 04:50:10 # Location: 72 Verbena Avenue, Dublin, County Dublin, D13 K6W6, Ireland.*

bar, and the vertically scrollable panel displaying the retrieved results in groups. Besides that, the system keeps three E-LifeSeeker modifications:

- **Ranked List Clustering**: Our system will cluster images based on temporal features and display top-3 with the highest scores.
- **Temporal Search**: LifeSeeker 6.0 has the ability to adjust the selected moment and its temporally-related images by controlling the temporal range between them.
- **Relevance Feedback**: The system recommends a question related to the visual content of users' desired images.

In order to offer users more options and improve their imagination, we add a detailed description for every image. The caption is CapMeta (3.3), which contains a detailed description as well as location and time, which will help users imagine better. Moreover, if the input query is identified as a question, the system will display the possible answer. For example, in the figure 4, with the input question as "What is sitting next to the guitar in the first month of 2020?", the interface returns the answer "teddy bear" for the corresponding image.

4 CONCLUSION

In this paper, we demonstrate the new adjustments for our lifelog retrieval system - LifeSeeker 6.0 at the 7th Lifelog Search Challenge. Firstly, we built a description corpus - CapMeta for the image collection which combined the generated description from the FUSECAP model and the metadata. Beyond the conventional image-text pair similarity, we implemented two streams of contrastive models for embedding, one for image and text, and the other for text and text. By doing so, we leveraged the linguistic aspect of our system in matching the query with the huge image collection. Moreover, we also added the description of CapMeta for every image and its possible answer for the QA task, which can support the users to interact with the system effortlessly. With these improvements, we hope to gain good results in the competition as well as open new ideas for the lifelog retrieval domain.

ACKNOWLEDGMENTS

This publication has emanated from research supported in part by research grants from Science Foundation Ireland under grant numbers SFI/13/RC/2106_P2, SFI/12/RC/2289_P2 and 18/CRT/6223, and co-funded by the European Regional Development Fund.

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*. 41–46.
[2] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2023. Voxento 4.0: A More Flexible Visualisation and Control for Lifelogs. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. 7–12.
[3] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9365–9374.
[4] Duc Tien Dang Nguyen Graham Healy Jakub Lokoc Liting Zhou Luca Rossetto Minh-Triet Tran Wolfgang Hürst Werner Bailer Klaus Schoeffmann Cathal Gurrin, Björn Þór Jónsson. 2023. Introduction to the Sixth Annual Lifelog Search Challenge, LSC'23. In *Proc. International Conference on Multimedia Retrieval (ICMR'23)* (Thessaloniki, Greece) *(ICMR '23)*. New York, NY, USA.
[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709 [cs.LG]
[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2818–2829.
[7] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. CERT: Contrastive Self-supervised Learning for Language Understanding. arXiv:2005.12766 [cs.CL]
[8] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).
[9] Cathal Gurrin, Liting Zhou, Graham Healy, Bailer. Werner, Duc-Tien Dang-Nguyen, Steve Hodges, Björn Þór Jónsson, Jakub Lokoč, Luca Rossetto, Minh-Triet Tran, and Klaus Schöffmann. 2024. Introduction to the Seventh Annual Lifelog Search Challenge, LSC'24. *International Conference on Multimedia Retrieval (ICMR'24)*. https://doi.org/10.1145/3652583.3658891
[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*. 47–52.
[11] Maria Tysse Hordvik, Julie Sophie Teilstad Østby, Manoj Kesavulu, Thao-Nhu Nguyen, Tu-Khiem Le, and Duc-Tien Dang-Nguyen. 2023. LifeLens: Transforming Lifelog Search with Innovative UX/UI Design. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. 1–6.
[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.
[13] Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Minh-Triet Tran, Liting Zhou, Pablo Redondo, Sinead Smyth, and Cathal Gurrin. 2019. Lifeseeker: Interactive lifelog search engine at lsc 2019. In *Proceedings of the ACM Workshop on Lifelog Search Challenge*. 37–40.
[14] Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Thanh-An Nguyen, Hai-Dang Nguyen, Liting Zhou, Graham Healy, and Cathal Gurrin. 2020. Lifeseeker 2.0: Interactive lifelog search engine at lsc 2020. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*. 57–62.
[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. https://arxiv.org/abs/2201.12086
[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
[17] 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*. 13–17.
[18] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Thanh Binh Nguyen, Graham Healy, Sinéad Smyth, Annalina Caputo, and Cathal Gurrin. 2022. LifeSeeker 4.0: An Interactive Lifelog Search Engine for LSC'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge*. 14–19.

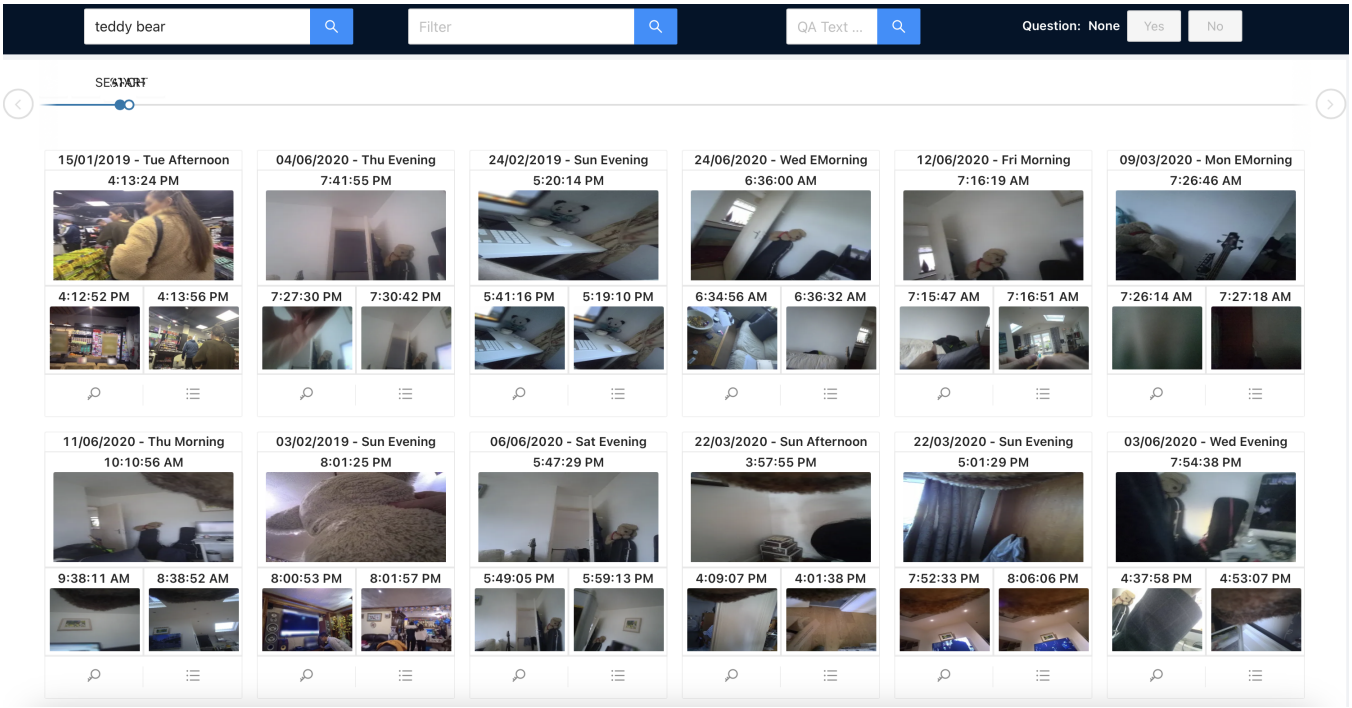

Figure 5: The User Interface of LifeSeeker 6.0.

[19] 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 *Proceedings of the 4th annual on lifelog search challenge*. 41–46.

[20] 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*. 59–64.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]

[24] Ricardo Ribeiro, Luísa Amaral, Wei Ye, Alina Trifan, António JR Neves, and Pedro Iglésias. 2023. MEMORIA: A Memory Enhancement and MOment RetrIeval Application for LSC 2023. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. 18–23.

[25] Joseph John Rocchio Jr. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing* (1971).

[26] Luca Rossetto, Oana Inel, Svenja Lange, Florian Ruosch, Ruijie Wang, and Abraham Bernstein. 2023. Multi-Mode Clustering for Graph-Based Lifelog Retrieval. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. 36–40.

[27] Noam Rotstein, David Bensaid, Shaked Brody, Roy Ganz, and Ron Kimmel. 2023. Fusecap: Leveraging large language models to fuse visual data into enriched image captions. *arXiv preprint arXiv:2305.17718* (2023).

[28] Klaus Schoeffmann. 2023. lifexplore at the lifelog search challenge 2023. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. 53–58.

[29] Florian Spiess, Ralph Gasser, Heiko Schuldt, and Luca Rossetto. 2023. The best of both worlds: Lifelog retrieval with a desktop-virtual reality hybrid system. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. 65–68.

[30] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[31] 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*. 24–29.

[32] 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*. 30–35.

[33] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7464–7475.

[34] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).