MyEachtra: Event-Based Interactive Lifelog Retrieval System for LSC'23

Ly-Duyen Tran
ly.tran2@mail.dcu.ie
Dublin City University
Ireland

Binh Nguyen
VNU HCM - University of Science
AISIA Research Lab
Vietnam

Liting Zhou
Insight Centre for Data Analytics, Dublin City University
Ireland

Cathal Gurrin
Dublin City University
Ireland

ABSTRACT

Retrieval is a fundamental challenge within the research community of lifelog and the Lifelog Search Challenge (LSC) has been an important annual benchmarking activity for interactive lifelog retrieval systems since 2018. This paper proposes MyEachtra (/maiAK-truh/), a system designed for the upcoming LSC'23 workshop. Improved upon MyScéal, which was the top performing system from LSC'20 to LSC'22, MyEachtra includes modifications to address the challenges of non-owner user understanding of lifelog contexts and open-ended lifelog question answering. Specifically, MyEachtra shifts the focus from images to events as retrieval units. Events are segmented using location metadata as well as visual and time differences between successive images. A pilot study on different approaches to aggregate images into events was conducted to test the automatic performance of the system, which showed promising results. For known-item queries, showing only the top 3 events proved to be adequate to find relevant images. However, future evaluation of the performance for ad-hoc and question-answering queries is necessary for a complete analysis of the MyEachtra.

KEYWORDS

lifelog, interactive retrieval system, pretrained models, user modeling

ACM Reference Format:
Ly-Duyen Tran, Binh Nguyen, Liting Zhou, and Cathal Gurrin. 2023. MyEachtra: Event-Based Interactive Lifelog Retrieval System for LSC'23. 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/10.1145/3592573.3593100

1 INTRODUCTION

Lifelogging refers to the process of passively capturing and storing personal data from everyday life activities in various formats such as 'point-of-view' photos (capture from wearable cameras), videos, location data, and biometrics data. With the popularity of wearable devices and digital technologies, more people, referred as lifeloggers, have started collecting their lifelogs. Lifelogging shows potential to bring many enhancements to the individual, such as enhanced well-being, higher levels of productivity as well as enhancing our understanding of personal experiences and behaviours.

Lifelog retrieval is one of the fundamental tasks that the lifelog community has been focusing on for the past few years. As more lifelog sources become readily available, lifelog research has gained increasing attention in many international workshops and activities such as Lifelog Search Challenges [7, 8, 24] and NTCIR Lifelog tasks [32]. Some challenges that are present in lifelog retrieval are: (i) the large amount of passively captured data requiring effective organisation and retrieval approaches, (ii) the lack of a single model for many modalities of lifelog archives, and (iii) an incomplete understanding of the user needs. The annual Lifelog Search Challenge provides a bench-marking opportunity for the research community to compare different approaches to lifelog retrieval, as well as introduce novel challenges to the community.

Known-item search has been the core task of the LSC since the first iteration in 2018, which requires finding one lifelog image within the provided dataset that is relevant to a query in natural language. To emulate the process of memory recall, the query consists of 6 hints which are gradually revealed every 30 seconds. In this kind of task, the main metric is precision and hit rate (as only one correct submission is needed). In 2022, two new tasks were introduced: ad-hoc and question-answering. Ad-hoc tasks accessed the recall of the submissions by asking for as many correct answers as possible. '*Find all the times I was buying whiskey in a store.*' is one example. On the other hand, question-answering (QA) tasks (e.g, '*What was the number of my office door (in 2019)?*') sought an image that identified the correct answer. This was, in a way, similar to the known-item search with the difference in the amount of given details. For all tasks, the score of each system is calculated based on the accuracy of the submissions as well as the submission speed.

The current iteration of the challenge, LSC'23 [4], modifies the QA task by requesting a free-from answer instead of an image. Research has been limited in this area. While a Lifelog Question Answering Dataset (LLQA)[22] was developed using the dataset from LSC'20[6], addressing the substantial jump from multiple-choice to open-ended questions will require significant further investigation.

Our previous system Myscéal [23], along with its updated versions Myscéal 2.0[26] and E-Myscéal[25], has scored highest in the last three editions of the LSC. The focus of this system has been

LSC '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0188-7/23/06.
https://doi.org/10.1145/3592573.3593100

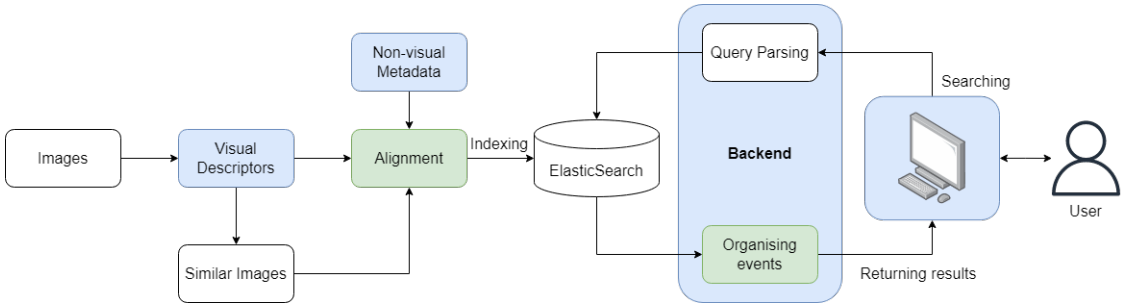

Figure 1: The MyEachtra overall pipeline follows the previous versions with slight modifications in blue modules and complete reformation in green modules to adjust to our event-based approach.

on providing an easy-to-use interface that minimises unnecessary interactions and speeds up the submission process. However, due to the COVID-19 pandemic, the evaluation benchmarks for the system in recent years were not able to include novice users who would not be familiar with lifelogging systems. As a result, the true performance of the system has not been fully tested, even though each version of the system was designed with novice users in mind. As a substitute for novice users, the expert performance from the previous year was analysed in preparation for the next event in 2023. Despite having more knowledge of lifelog retrieval than novice users, the expert still struggled to determine the relevance of some images, and there was a high risk of missing correct images. This inspires us to adopt an **event-based approach** to lifelog retrieval. We believe this will help users understand the context of the retrieved results more quickly and accurately.

In this paper, the proposed system, **MyEachtra** (/mai-AK-truh/), builds on the previous version (E-Myscéal) and includes modifications which are summarised as follows: (i) an event segmentation process influenced by location metadata; (ii) an event-based approach that exploits pretrained image-text cross-embedding models to develop representative embeddings for event; (iii) a redesigned user interface showing events and highlighting important images with relevant events; and (iv) a pipeline to apply video QA models on retrieved events to answer open-ended questions.

2 RELATED WORK

2.1 Lifelog Retrieval

Lifelog retrieval refers to the process of searching for and retrieving information from lifelogs. It typically involves customised search engines, which employ various ranking techniques to present the most relevant lifelog information in an understandable format to the user. As the volume and diversity of lifelog archives grow, efficient lifelog retrieval is crucial to allow the lifeloggers to easily access and review their past experiences.

As a preliminary attempt to address lifelog retrieval, the solution to the workshop tasks are in the form of individual images. Although there are some attempts with aggregating images into 'moments' (although moments are not yet clearly defined) [9, 15, 25], most systems—even the best performing ones—treat each image separately and turn the task into image retrieval with boolean filters for other modalities (such as map filtering for location, faceted filters for time, etc.). Traditionally, lifelog search systems annotate each lifelog image with textual information extracted from various computer vision models and the ranked list result is generated by comparing the 'keywords' in the queries with the annotated concepts. Various teams that participated in the past LSC workshops employed such an approach [5, 18, 19, 23, 26]. This approach results in acceptable performance but faces difficulties when the query becomes more complicated. One approach has been to employ more models in preprocessing to cover more contexts that are potentially searched for. However, with the increasing popularity of cross-modal embeddings, such as CLIP [17], ALIGN [10], and CoCa [30], many teams have started adapting these pre-trained models into their systems[1, 15, 25]. Embedding-based models provide a greater semantic understanding of images compared to disjointed concepts and have proved their usefulness in recent workshops. Nevertheless, these new methods still perform search on the image-level, ignoring the fact that life experiences are temporal in nature and many topics in recent LSC editions are built upon that temporal relationship of lifelogs (e.g repeatedly *reading the manual instruction* while *building a computer*, then *going for a coffee*).

The main motivation of this work stems from a speculation that, these systems could be more robust if they take into consideration the temporal nature of lifelogs. By extending the semantic understanding of embedding-based models from the image-level to a longer period of time, a lifelog system could perform more complicated searches.

2.2 Adapting Image-Text Models for Video Modelling

As mentioned in previous section, image-text embedding models have gained significant popularity. The idea behind both models is training a text encoder and an image encoder with contrastive loss on a large number of image-text pairs. Over the past year, there have been numerous attempts to inherit the knowledge from such image-text models to model videos with minimal re-training. One approach is using late fusion module to aggregate video frame features. Examples include using mean pooling or attentional poolers [28, 30], Transformer Encoders[11], calculating a weighted mean of frame-wise embeddings[3], or using K-means to return K representative embeddings, of which the one that has the maximum similarity score with a query will be registered during retrieval[16].

Other works proposed novel visual models which are initialised with pretrained weights from a image-text model. CLIP2CLIP[14] modifies the linear projection layer of the Vision Transformer (ViT) to take into account the time axis of videos. To further improve the efficiency of the model, CenterCLIP[31] suggests clustering video patches before passing them into the ViT model.

Leveraging image-text models can also be used in various downstream tasks including video QA. FrozenBiLM [29] utilises frozen pretrained visual encoders by integrating lightweight adapter modules to enable zero-shot video QA. VideoCoCa [28] improves on image CoCa and reuses attention poolers that are parts of the pretrained image-text model without further retraining.

Inspired by these works, we choose to employ pre-trained image-text embedding models to model lifelog events. As the Lifelog Search Challenge involves processing a large number of images within a short period, we opt for text-independent methods for computing event embeddings offline. This enables us to process each query more quickly at search time.

3 OVERVIEW OF THE MYSCÉAL SYSTEM

In this section, we briefly introduce the previous Myscéal system, which MyEachtra is based on. The pipeline of Myscéal has remained mostly unchanged since the first version, illustrated in Figure 1.

The data processing components provided visual descriptors and non-visual metadata, such as GPS coordinates, semantic location names, time, and date, for each image. Similar images were previously computed at this stage using VGG16 [21] and SIFT[12, 13] features. The Alignment module took into account different sources of processed data and create documents with image-based keys. All documents are then indexed in ElasticSearch [1] to enable high-speed searches.

User interactions were performed on a desktop interface, as seen in Figure 2. The backend system was implemented using Django[2] to support communication between the user interface and ElasticSearch. Myscéal aimed to support a novice user by minimising interaction steps, so it used a full-text search instead of relying on a faceted filter panel. The necessary information, both non-visual metadata and visual descriptors, could be directly parsed from the textual query by a custom query interpreter.

Myscéal also supported searching for multiple queries based on their temporal relationships, which is a distinguishing characteristic of lifelogs. This is shown by the design of three separate query boxes ('before', 'during', and 'after') at the top of the interface, and reinforced by showing the search results in triplets, putting the images in their temporal context.

In some cases, location information could not be well represented in text, and the user could use the map panel on the top right to locate the target location and perform filtering. Additionally, the saved section allowed the user to put aside images that they are not yet certain about for later consideration.

Another important aspect of Myscéal was the Event View displayed, which was designed to help the user understand the context around an image by showing where the image sat on a timeline covering the entire lifelog.

[1] https://www.elastic.co/
[2] https://www.djangoproject.com/

4 INTRODUCING MYEACHTRA

4.1 Data Processing

Extra metadata We kept the data processing steps from Myscéal mentioned in previous section with two changes. First, we simplify the data processing components by using CLIP embeddings directly to determine image similarities. Specifically, CLIP-H/14 is used, which was trained with the LAION-2B English subset of LAION-5B [20] using OpenCLIP [17]. Moreover, we incorporate an addition source of metadata: semantic names, which were produced by VAISL[27]—a GPS processing method. Each semantic location also comes with the location type (e.g. Korean restaurant, library, etc.) to provide extra information when the user needs. However, as a large amount of GPS data was missing from the original data, the semantic locations identified sometimes may not be correct. However, this information helps tremendously with segmenting lifelog, which is discussed next.

Event segmentation Moving to an event-based approach, it has become increasingly necessary to have a well segmented lifelog. Thus, we define event boundaries based on (i) the the change of semantic locations acquired from VAISL, (ii) cosine distance between successive images, and (iii) the time gap between two images. To further assist the use of date filters in the query, we also split the event if they occur over two days (e.g, staying at home overnight). Using a cosine distance threshold of 0.3 results in 167,570 separate events.

4.2 Event-Based Approach

The main enhancement for MyEachtra is that, instead of comparing each image in the dataset to the query using cosine distances, we compare *events*. We illustrate how to turn image embeddings into event embeddings. By using CLIP, we denote the pre-trained encoders as $\omega(u) = \boldsymbol{w}$ and $\theta(t) = \boldsymbol{c_t}$ which encode image u and text t into $\boldsymbol{w}, \boldsymbol{c_t} \in \mathbb{R}^d$. Assume an event e is composed of s images such that $e = u_1, u_2, ..., u_s$. Therefore, we can join the embedding of each image into a matrix $Z = [\omega(e) = \boldsymbol{z_1}, \boldsymbol{z_2}, ..., \boldsymbol{z_s}]$, where $Z \in \mathbb{R}^{d \times s}$. We need an aggregation function Λ that maps $Z \in \mathbb{R}^{d \times s}$ into an global event representation $\boldsymbol{c_e} \in \mathbb{R}^d$. In the context of the LSC, Λ is preferably independent from the query t. Several options are possible from previous work in video-text models.

Mean Pooling A simple yet effective way of combining a list of embeddings is average pooling over the temporal dimension. Mean pooling is often used as a baseline to compare new video models.

Clustering Portillo et al.[16] experimented on clustering the events and selected the cluster centres as representative embeddings. The only modification from their method is that instead of using K-means clustering method, we employ OPTICS[2] to address the vastly varied lengths of events.

Transformer encoders The most popular technique for temporal modelling in videos (as well as events in this system) is to use transformer encoders and learn a self-attention mechanism to emphasise important images. Note that since the outputs of transformer encoders are still in a sequential format, they are average pooled to create the global embedding.

Weighted Mean Another way to work with these outputs is passing them through a Linear Layer (where the output dimension

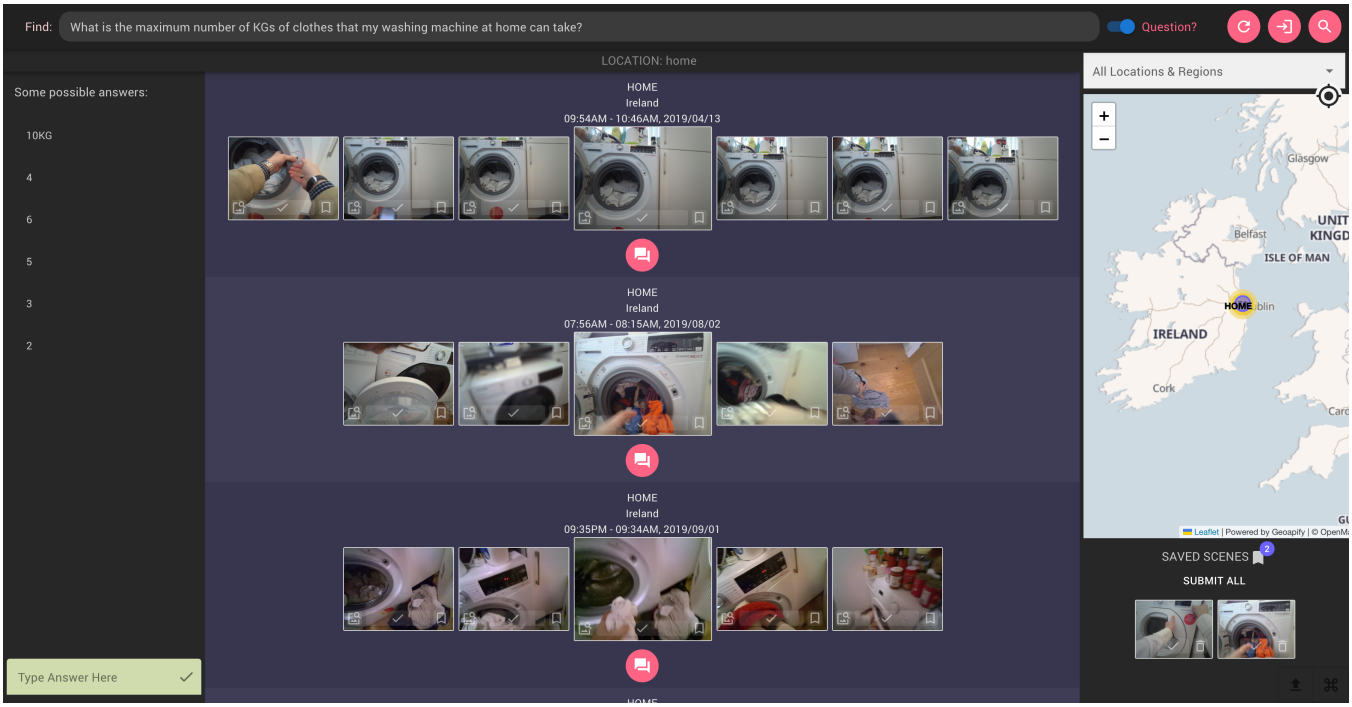

Figure 2: MyEachtra user interface. For non-QA tasks, the left panel is hidden.

is 1) to produce images weights, indicating how important an image is in the event, as described in [3].

After getting the event embeddings, we continue to use the cosine similarity for the retrieval process. The similarity score is summed with other scores (TF-IDF for location names, GPS filters, time filters, etc.) within ElasticSearch, similarly to the previous year. We refer to these as the *event scores* in later subsection. In Section 5, we evaluate the results using the previously mentioned options.

4.3 Displaying Events

We redesigned the user interface to show the resulting ranked events (rather than images) in a way that is easy to understand and highlights relevant information such as location, time, and highly ranked images within an event. After getting the ranked results from ElasticSearch, to further reduce repetitive information, if there is location changes between some events, we merge them together as one *row*. The reason behind this is that our event segmentation still does not account for longer events and that some activities in one location can belong to different event segments.

To find the best images within each *row* to display, we first calculate the cosine similarities between each image and the user's query. Each *similarity score* is then multiplied with the *event score* returned by Elasticsearch (as mentioned in last section). Finally, we apply the softmax function over the scores to get the *image scores*. The softmax function can help emphasise the most relevant images and reduce the impact of outliers. We limit this calculation to the top-100 resulting events and repeat the process when the user requests more results.

When aggregating event images by using Weighted Mean, we also get the image weights as the output of the model. In that case, these weights are multiplied with the cosine similarities and event scores before the softmax function is applied.

To help the user quickly identify the most highly scored image within each group, we highlight it and place it in the middle of the row. This design choice was made to draw the user's attention to the most important visual information and reduce the need for extra scanning and searching. In addition, up to the next six most relevant images (three on each side, left and right) are also shown to the user as they are not only most likely relevant but also provide additional context to improve the user's understanding of the event. An example could be seen in Figure 2.

Regarding known-item queries, the user can submit an image by clicking on a checked button displayed on it. On the other hand, for ad-hoc queries, users can submit the entire event by holding down the Shift key and clicking on any image's submit button.

4.4 Question Answering

To generate the list of answers, we utilise FrozenBiLM [29], a video question-answering (QA) model, on the most relevant scenes for the user's query. The model's architecture consists of a *frozen* bidirection language model (BiLM) and a *frozen* pretrained visual encoder (CLIP VIT-L/14) connected by adapter modules. These adapter modules are multilayer perceptrons with residual connection and are insearted after each self-attention layer and each feed-forward layer of the language BiLM's model. To adapt FrozenBiLM to open-ended videoQA (or lifelog QA in this system), a prompt is created as follows: '[CLS] Question: <Question>? Answer: [MASK].'. The

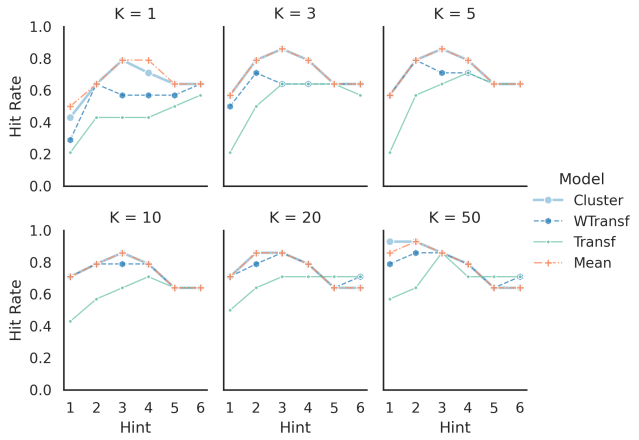

Figure 3: Hit Rate at K at different hints on four approaches.

limitation of this model is that it is essentially a classification model that is only capable of choosing the best answers from a fixed set.

In MyEachtra, the question is used directly as a search query to find the most relevant events. The top-10 events are passed through FrozenBiLM to get potential answers, which are shown on the left panel of the user interface in Figure 2. Nonetheless, users can choose to run the model again for any event they find interesting by clicking on the QA button underneath each row. Once an answer is found, users have the option to swiftly copy it by clicking on it and then pasting it into the submission input field located at the bottom left corner. Alternatively, if they can deduce the answer from the displayed events, they can manually type it into the box and submit it.

5 EVALUATION USING LSC'22 QUERIES

We carried out an automatic evaluation to assess the performance of our event-based approach. 14 known-item queries of the LSC'22 queries were used in this pilot study. Similarly to last year's experiment, they are split into 'before', 'main', and 'after' hints before requesting the ranked list from the backend system. The results are measured using Hit Rate at K ($H@K$), with an adjustment for events. $H@K$ here means that one of the target images are included in the first K event (whose a maximum of 7 images are shown in the user interface). This metric could provide a baseline for the system's performance that may be comparable to the performance of a novice user because it ignores complex user interactions that the system supports.

We experiment on four different approaches to aggregate image features and the results are shown in Figure 3. The configurations of each experiment are as follows:

- **Mean**: the mean pooled embedding are used as the global event embedding. No training is required.
- **Cluster**: OPTICS clustering is applied for each scene with `max_eps=0.5` and `min_samples=2`

Table 1: Mean $H@K$ for LSC'22 queries using Mean Pooling. We are most interested in the modified version of H@3 because (i) once the user find the correct answer, more hints are not needed and (ii) the user interface can handle three events at a time.

Hint	H@1	H@3	H@5	H@10	H@20	H@50	Mod H@3
1	0.50	0.57	0.57	0.71	0.71	0.86	0.57
2	0.64	0.79	0.79	0.79	**0.86**	**0.93**	0.79
3	**0.79**	**0.86**	**0.86**	**0.86**	**0.8**6	0.86	0.86
4	**0.79**	0.79	0.79	0.79	0.79	0.79	0.86
5	0.64	0.64	0.64	0.64	0.64	0.64	0.86
6	0.64	0.64	0.64	0.64	0.64	0.64	0.86

- **Transf**: we trained *one* layer of PyTorch [3] implementation of Transformer Encoders for 10 epochs using the captions described in LLQA dataset [22] with `n_head=8` and `d_model=1024`. The outputs are mean pooled.
- **WTransf**: same settings with Transf. The outputs are used to created a weighted mean embedding.

Surprisingly, the straightforward method of mean pooling achieves the highest results in most cases. As for clustering, not only the search space has increased, the hit rates at $K = 1$ are also slightly lower. Furthermore, despite having more parameters, both the weighted mean (WTransf) and averaging the output (Transf) from Transformer encoders produce generally worse performance, especially at lower values of K and when fewer hints are used. This could be explained by the limited size of the training dataset, which contains of only 13,317 captions with low variety.

The best performing setting is recorded in Table 1. From the experiments, we observe that more hints do not mean better results. In fact, the system seems to perform the best when 2-3 hints are given, without 'before', 'after', or misleading hints (e.g wrong year). Thus, we also report a modified H@K metric, denoted as Mod H@K in the table, where $H@K(i) = max(H@K(i), H@K(i-1))$ to account for the fact that more searches are not needed after the correct submission has been made. Furthermore, we are also aware of the trade-offs when choosing to show events instead of individual when it comes to the amount of results that can be effectively be displayed on the user interface. Our user event-based interface currently can fit 3 events at most, thus we are most interested in Mod H@3, when we assume *no scrolling is needed*. Here we could see that the answer for 57% (8 out of 14) of the queries can be found using only the first hints. With more hints, the user can find the correct image of 86% of the queries (12 out of 14).

6 CONCLUSION

In this paper, we present modifications to Myscéal that shift the focus of lifelog retrieval from images to events, aiming to move towards a unified model for lifelog archives. Our new system, MyEachtraexploits pretrained image-text models to create event embeddings. We conducted pilot experiments and found that averaging image embeddings to create event embeddings is the most suitable

[3] https://pytorch.org/

approach for MyEachtra at this stage, resulting in a reduced search space without sacrificing performance. Additionally, we adjust the user interface to show relevant events and focuses on contextual information effectively. However, there is more room for improvements regarding the performance of FrozenBiLM on lifelog data as the answers are still not directly found from the model's outputs. In future work, we plan to explore more advanced techniques to aggregate image embeddings into event embeddings and incorporate more diverse datasets for training. User studies are also necessary to gain more knowledge for the future systems.

ACKNOWLEDGMENTS

This publication has emanated from research supported in party by research grants from Science Foundation Ireland under grant numbers SFI/12/RC/2289, SFI/13/RC/2106, 18/CRT/6223 and 18/CRT/6224.

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 on Lifelog Search Challenge*. 2–7.

[2] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record* 28, 2 (1999), 49–60.

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2022. A CLIP-Hitchhiker's Guide to Long Video Retrieval. *arXiv preprint arXiv:2205.08508* (2022).

[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] Aaron Duane, Cathal Gurrin, and Wolfgang Huerst. 2018. Virtual reality lifelog explorer: lifelog search challenge at ACM ICMR 2018. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*. 20–23.

[6] Cathal Gurrin, Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Björn Þór Jónsson, Jakub Lokoč, Wolfgang Hurst, Minh-Triet Tran, and Klaus Schoeffmann. 2020. An Introduction to the Third Annual Lifelog Search Challenge, LSC'20. In *ICMR '20, The 2020 International Conference on Multimedia Retrieval*. ACM, Dublin, Ireland.

[7] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Dang Nguyen, Duc Tien, Michael Riegler, Luca Piras, et al. 2019. Comparing approaches to interactive lifelog search at the lifelog search challenge (LSC2018). *ITE Transactions on Media Technology and Applications* 7, 2 (2019), 46–59.

[8] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2022. Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *Proc. International Conference on Multimedia Retrieval (ICMR'22)*. ACM, Newark, NJ.

[9] 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 on Lifelog Search Challenge*. 20–26.

[10] 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.

[11] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*. Springer, 105–124.

[12] D. G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 2. 1150–1157 vol.2.

[13] David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60, 2 (Nov 2004), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

[14] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval. *arXiv preprint arXiv:2104.08860* (2021).

[15] 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.

[16] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. 2021. A straightforward framework for video retrieval using clip. In *Pattern Recognition: 13th Mexican Conference, MCPR 2021, Mexico City, Mexico, June 23–26, 2021, Proceedings*. Springer, 3–12.

[17] 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.

[18] Ricardo Ribiero, Alina Trifan, and Antonio JR Neves. 2022. MEMORIA: A Memory Enhancement and MOment RetrIeval Application for LSC 2022. In *Proceedings of the 5th Annual on Lifelog Search Challenge*. 8–13.

[19] Luca Rossetto, Ralph Gasser, Silvan Heller, Mahnaz Amiri Parian, and Heiko Schuldt. 2019. Retrieval of structured and unstructured data with vitrivr. In *Proceedings of the ACM Workshop on Lifelog Search Challenge*. 27–31.

[20] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=M3Y74vmsMcY

[21] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[22] Ly-Duyen Tran, Thanh Cong Ho, Lan Anh Pham, Binh Nguyen, Cathal Gurrin, and Liting Zhou. 2022. LLQA-Lifelog Question Answering Dataset. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part I*. Springer, 217–228.

[23] Ly-Duyen Tran, Manh-Duy Nguyen, Nguyen Thanh Binh, Hyowon Lee, and Cathal Gurrin. 2020. Myscéal: An Experimental Interactive Lifelog Retrieval System for LSC'20. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*. 23–28.

[24] Ly-Duyen Tran, Manh-Duy Nguyen, Duc-Tien Dang-Nguyen, Silvan Heller, Florian Spiess, Jakub Lokoč, Ladislav Peška, Thao-Nhu Nguyen, Omar Shahbaz Khan, Aaron Duane, et al. 2023. Comparing Interactive Retrieval Approaches at the Lifelog Search Challenge 2021. *IEEE Access* (2023).

[25] 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 on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22)*. Association for Computing Machinery, New York, NY, USA, 32–37. https://doi.org/10.1145/3512729.3533012

[26] 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 *Proceedings of the 4th Annual on Lifelog Search Challenge*. 11–16.

[27] 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. in press.

[28] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. 2022. Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners. *arXiv preprint arXiv:2212.04979* (2022).

[29] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155* (2022).

[30] 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).

[31] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2022. CenterCLIP: Token Clustering for Efficient Text-Video Retrieval. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 11–15, 2022, Madrid, Spain*.

[32] Liting Zhou, Cathal Gurrin, Graham Healy, Hideo Joho, Thanh-Binh Nguyen, Rami Albatal, Frank Hopfgartner, and Duc-Tien Dang-Nguyen. 2022. Overview of the NTCIR-16 Lifelog-4 Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*. National Institute of Informatics, 130–135.