An Experiment in Interactive Retrieval for the Lifelog Moment Retrieval Task at ImageCLEFlifelog2020

Ly-Duyen Tran[1], Manh-Duy Nguyen[1],
Binh T. Nguyen[2,3,4], and Cathal Gurrin[1]

[1] Dublin City University, Dublin, Ireland
[2] AISIA Research Lab
[3] University of Science, Ho Chi Minh City, Vietnam
[4] Vietnam National University, Ho Chi Minh City, Vietnam

Abstract. The development of technology has led to an increase in mobile devices' use to keep track of individual daily activities, as known as Lifelogging. Lifelogging has raised many research challenges, one of which is how to retrieve a specific moment in response to a user's information need. This paper presents an efficient interactive search engine for large multimodal lifelog data which is evaluated in the ImageCLEFlifelog2020 Lifelog Moment Retrieval task (LMRT). The system is the modified version of the Myscéal demonstrator used in the Lifelog Search Challenge 2020, with the addition of visual similarity and a new method of visualising results. In interactive experimentation, our system achieved an F1@10 score of 0.48 in the official submission but can be significantly improved by implementing a number of post-processing steps.

1 Introduction

As defined in [8], lifelogging refers to the process of using technology to keep a log of one's daily activities using various media, such as images, location, or biometrics. The stored data is called a lifelog and can help its owners understand their activities and recall some memorable moments in their lives. In order to be useful to the individual, a lifelog should have some form of retrieval tool that can assist them in seeking remembered information. Many collaborative benchmarking fora have been started to assist the research community to make progress, buy defining research challenges and releasing test collections. For example, the NTCIR Lifelog task (from 2015-2019) [5], the Lifelog Search Challenge (LSC) [7] since 2018 and the ImageCLEFlifelog [15] are all examples of such fora. Given the volumes of lifelog data that an individual can generate, any retrieval system needs to provide accurate retrieval facilities in a timely manner. Additionally, such a tool should have a user-friendly design that can help users to operate

efficiently. The ImageCLEFlifelog2020 Lifelog Moment Retrieval task (LMRT) [15], which is one of four main tasks in the ImageCLEF2020 campaign [9], requires participating interactive retrieval systems to retrieve all possible images matching given topics and do this within a time-limit of five minutes per topic.

In this paper, we address this ad-hoc interactive retrieval challenge for lifelogs by enhancing the performance of our pre-existing lifelog retrieval system Myscéal [18]. Since the challenge of LMRT is to find all the moments of interest that match the information need, we implemented visual similarity matching, to assist the user to find visually related content to one positive example. We also added an extra faceted window in the user interface to show a detailed summary of the retrieved results and to provide the user with a result filtering mechanism. Additionally, the ranked result display area has been adjusted to be suitable for the LMRT objectives. Consequently, the contribution of this paper is twofold; firstly in describing the enhanced version of Myscéal, and secondly in describing the result of an interactive retrieval experiment to evaluate the performance of the new system with three types of users; an expert user (the system developer) who is familiar with the tool, a knowledgeable user (the owner) of the dataset who is a novice user, and a full novice user who does not know the tool or the dataset. Finally, we report on the automatic offline post-processing steps and show that they can improve the scoring metrics significantly.

2 Related Work

There have been a number of interactive lifelog retrieval systems described in the literature. One of the pioneers in this field, Doherty et. al.[4] supported similarity search through event archives by using a simple color-based descriptor. Meanwhile, LEMoRe [16], being an example of a more recent system, introduced the concept of integrating images with their semantic context and applied natural language processing to handle textual queries. The top three systems in ImageCLEFlifelog 2019 LMRT [2] last year also followed a similar idea. The HCMUS team [11] annotated each image with its concepts along with the inferred colour of detected objects. Moreover, they extended the semantic concepts with the lifelogger's habits and activities. The BIDAL team [3] used the concepts and other textual information to generate the Bag-of-Word (BOW) vectors representing each image. They then combined the BOW vector generated from the query and used them to find suitable images. In contrast, The ZJUTCVR team [20] viewed the challenge as a classification problem with additional pre-processing steps to remove the blurred images.

Our prior system at the LSC'20 [6], Myscéal [18] approached the issue as a traditional text search challenge, in which visual concept annotations and various forms of metadata were indexed as if conventional text. For this task, we implemented two new features (described in the next section) and removed some functions that we think not useful for this LMRT challenge. We also evaluated the system by designing the experiments with three users representing three different

usage scenarios. Finally, the post-processing steps were applied to achieve a higher score specifically for the challenge scoring mechanism.

3 Our Interactive Retrieval System

3.1 System Overview

The modified Myscéal retained the processing pipeline of the original version as depicted in Figure 1, which follows a typical structure for a lifelog search engine as introduced in [19]. The visual concepts of each image were initially derived from the given metadata and augmented with the output of the object detector from DeepLabv3+ [1]. Those annotations, along with other information such as locations and time, were then indexed in the open-source search engine (ElasticSearch). The input query was analysed to extract the primary information and enlarged by our expansion mechanism (see section 3.4), then matched with our inverted index to find the potentially relevant images that were ordered by the ranking function and presented to the user. The readers are referred to the original work in [18], which describes in detail how this process operates.

In this version, besides the concepts detected from the descriptor, we employed the Microsoft Computer Vision API[1] service to enrich the visual annotations further. To provide an optimized interactive retrieval system for LMRT, we introduced three updates to the previous system; visual similarity, user interface, and summary panel. We will now describe each of these components.

3.2 Visual Similarity

The Myscéal system was developed for the LSC'20 challenge, which required participants to find any single relevant image for a given topic as fast as possible. The ImageCLEFlifelog2020 LMRT task is different in that it seeks a list of suitable images for a given query. This is a subtle difference that requires a different retrieval engine. Firstly we implemented a visual similarity feature to facilitate the user in finding all visually similar images to any given image. We measured the similarities between images by using the cosine distance of their visual features, which comprised SIFT [12,13,14] and VGG16 [17] features. We did not include visual concept descriptions because the intention of visual similarities is to provide users a different way of searching that is independent as possible to the text-based retrieval. To ensure real-time retrieval, we made this process offline and indexed a list of similar images of each image in the ElasticSearch engine prior to accepting any user queries.

3.3 User Interface

The LSC'20 challenge included topics that had a definite temporal element (e.g. going to the cafe after work), which was not expected in the LMRT topics, hence

[1] `https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/`

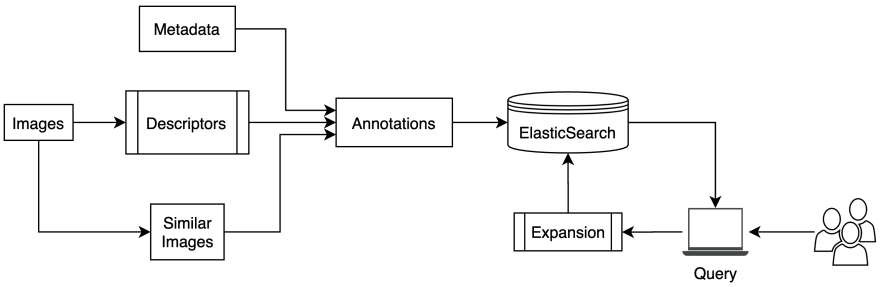


Fig. 1: Our system's pipeline included the offline data indexing stage and the online retrieval stage. ElasticSearch was used to store the annotations including detected objects and visual similar photos extracted from every image and its metadata. In the retrieval phase, a query given by users is firstly expanded to include more information, then is compared with the data in the storage and shown to users.

we replaced the temporal query mechanism of Myscéal with a more conventional search box for users to enter the query. Whenever the user submits the query, the retrieval results are shown in the middle of the screen in a ranked list, supporting both vertical and horizontal scrolling. Each row contains concatenated grouped images, which are in decreasing order to the similarity to the query and grouped by day. Each group represents a moment or an activity in which there are many visually similar images. This structure not only reduces the visual and cognitive load on a user, but it also allows for a higher level of recall. This is because the user will have a clearer view with less identical images leading to have more time and high chances to spot out relevant photos in the ranked list.

Clicking on an image in the ranked list opens a detail window containing all images taken on the same day in an event-level view. This event view, as illustrated in 3, is arranged with three parts to show the hierarchy of event grouping. Visually similar images (in the first row) are grouped together and their thumbnail is shown as one item on the second row; the bottom row indicates the broadest level of grouping: each image in this row is a thumbnail of an activity that happens in the same location (walking in the airport, sitting inside the airplane). Using this, the user can browse the images through the day at a faster pace.

There is also a small visual similarity icon at the bottom of each image that allows users to open the similarity window listing all similar photos of the selected image. Every image in this window contains a button that opens the similarity panel.

On the top-right pane of the user interface, we show the "Saved scenes" section. Whenever users find a relevant image, they can save that photo quickly with the help of the saving icon appearing at the bottom of every image. The bottom-right map panel, which remains the same as the previous version, works as a location filter, or illustrates the location of an image.

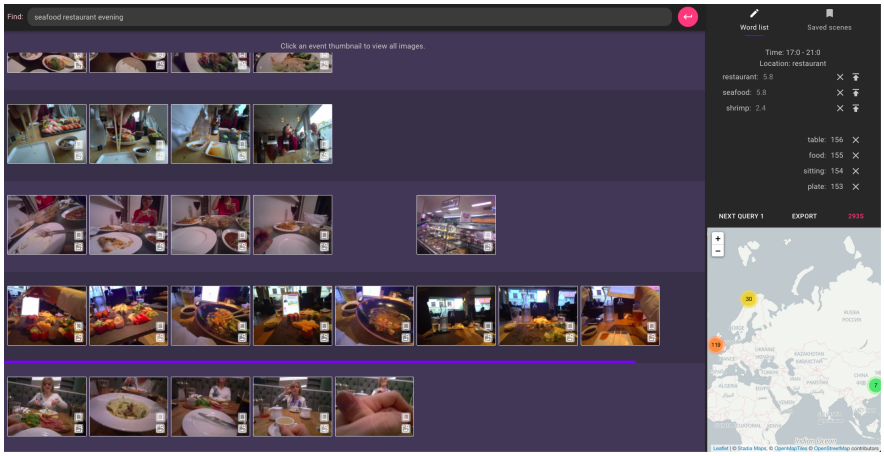


Fig. 2: Our main user interface with the search bar on the top and the main panel below for showing the result. The top-right area presents the "Word list" panel while the map indicating the geospatial location of photos is at the bottom-right area.

Additionally, we introduced a new "Word List" panel appearing in the same area with the "Saved scenes" panel. This feature will list all concepts used in the retrieval system and their scores so that users can adjust (e.g., increase the scores, or remove the concepts) to have better results.

3.4 Word List Panel

The word list and their corresponding scores comprise the query expansion process, as introduced in [18]. We employed heuristic rules to assign scores to each word in the expansion list: (1) A word will have the score of α if it is an exact match of the user query; (2) if the word is the result of the synonym, hypernym or hyponym expansion, we assign its score β; (3) if the word is the result of the Word2Vec similarity expansion, assign the similarity score; and (4) if the word is the result of multiple expansion (i.e., it is similar to several words), the final score will be the highest one. Furthermore, to reduce the workload on the search engine and not confuse users by showing a long list of words, only 20 highest scoring words were chosen. We set $\alpha = 3, \beta = 0.95$ after some empirical trials.

In addition to the scoring word lists, we also provided a quick summary of the results by displaying the most frequently occurring visual concepts. This way, users can choose to remove some of the irrelevant contexts. For example, searching for *kitchen* might result in a distracting amount of images that show a *window*. Meanwhile, the user wants to look for something from another point of view. Using the summary feature, the user can remove the window to have a more focused result view.

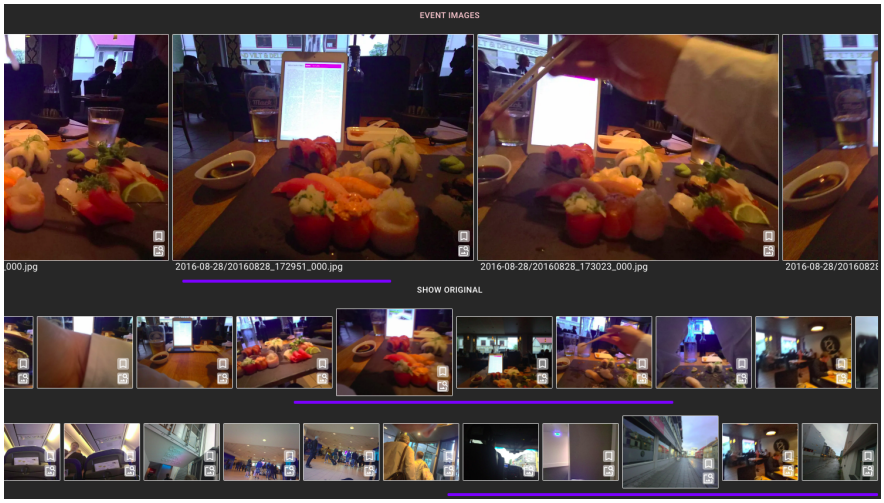


Fig. 3: Events view windows

4 User Experiments Setup

To evaluate the system's performance, we designed a user experiment by asking three users, the expert system creator, the knowledgeable data owner, and a novice user, representing three use-cases. The developer would know clearly and could operate the system in the best way, while the data owner, or lifelogger, was expected to know the answers to all topics. In contrast, the novice user was not familiar with either the system or the dataset. All users had a maximum of 5 minutes to solve each query, which did not include the time for reading the topic. Prior to the experiment, all users were given the opportunity to familiarise themselves with the system by processing some sample topics under the guidance of the system creators. After finishing each query, users were required to record their searching time, displayed on the system, the distance of mouse movement on the screen (measured in meters), the number of mouse clicks, and the number of keystrokes (all gathered using a third-party tool). The three users were encouraged to utilize the entire 5 minutes to retrieve images from as many relevant events as possible.

5 Results

5.1 LMRT Task

Evaluation of system performance (user performance) is via three metrics suggested by the LMRT organizers; Precision at 10 (P@10), Cluster Recall at 10 (CR@10), and the F1 Score at 10 (F1@10). Since the users were instructed to focus on finding images from all relevant events instead of selecting all images in

a single event, the post-processing steps are necessary to improve the recall score in LMRT. For each selected image, we expanded the result by finding all similar photos within a given temporal window to augment the submitted results. After performing the expansion for all images, we rearranged the result by moving the user-selected images in each cluster to the top rank to ensure a higher P@10 score. All of the steps described above were processed automatically offline after the users finished their search sessions. The $F1@10$ scores [15] for each user are shown in the Table 1. Please note that we present both the raw user runs and the expanded post-processed runs separately and that the official runs were the raw-runs.

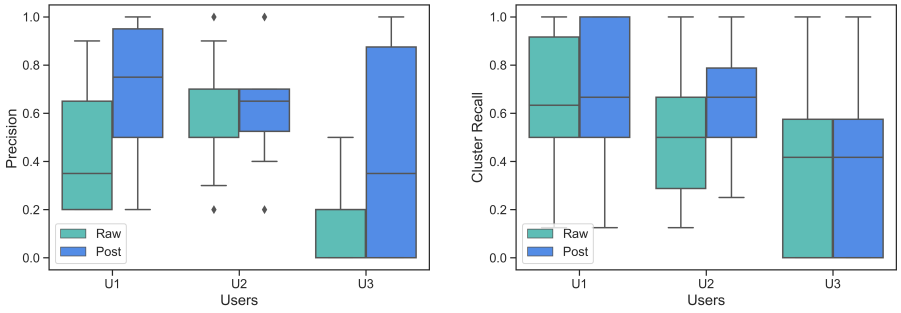
Table 1: $F1@10$ scores of three users (U1: Expert, U2: Owner, U3: Novice). The Raw column indicates the original result from users and the Post column shows the score after applying post-processing steps. The *dot* means the user could not find the answer for that question. Numbers with the $*$ are the highest number in that topic.

Query	**U1 (expert)**		**U2 (owner)**		**U3 (novice)**	
	Raw	Post	Raw	Post	Raw	Post
1	0.33	1^*	0.5	0.58	0.28	0.67
2	0.22	0.22	0.22	0.72^*	.	.
3	0.33	0.57	0.94	1^*	0.33	1^*
4	0.22	0.22	0.27	0.31^*	.	.
5	0.68	0.68^*	0.68	0.68^*	.	.
6	0.5	0.5^*	0.25	0.25	0.25	0.25
7	0.68	0.89^*	0.44	0.69	0.4	0.69
8	0.33	1^*	0.33	0.75	.	.
9	0.31	0.73	0.68	0.8^*	0.31	0.77
10	0.5	0.5	0.5	0.5	0.5	0.5
Overall	0.41	0.63^*	0.48	0.63^*	0.21	0.39

The experimental results indicate that the lifelogger, who knows best about the dataset, got the highest overall score among all users in both the original and the refined answers. The authors of the system also obtained a comparable score and even achieved the same as the data owner after the post-processing step. The average $F1@10$ score of the novice user was lower than the others with 0.21 and 0.39 for the raw and the modified result, respectively. One possible reason is this user was not successful in solving nearly half of the tasks in the challenge. The post-processing stage had a significant impact on the score as it improved the overall value by at least 30%. It could boost the original to the absolute value of 1 as two steps were expected to capture full precision and recall criterion. There were, however, some queries that remained the same after the second stage. Another interesting detail was that the system creator had four

topics whose scores were higher than others, while that of the lifelogger was just three.

The detailed distribution of the precision and recall of each user is illustrated in Figure 4. There was a significant improvement in the precision score of the developer and the novice user. It was because they only submitted a few images within a cluster to spend more time discovering more groups leading to not enough answers for the evaluation, hence getting the low P@10 in the initial submission. The retrieving behaviors of these two users also allowed them to found out many distinct events and ranked them at the top places in the submitted results hence gained sufficient CR@10 scores. Therefore the refining steps almost had no impact on this metric and the CR@10 of both volunteers slightly remained. In contrast, the lifelogger tended to select all relevant images he saw on the screen making the answers from different clusters were not in the top 10 anymore. The post-processing algorithm could fix this issue by rearranging the results and improving the scores shown in Figure 4b.



(a) P@10 (b) CR@10

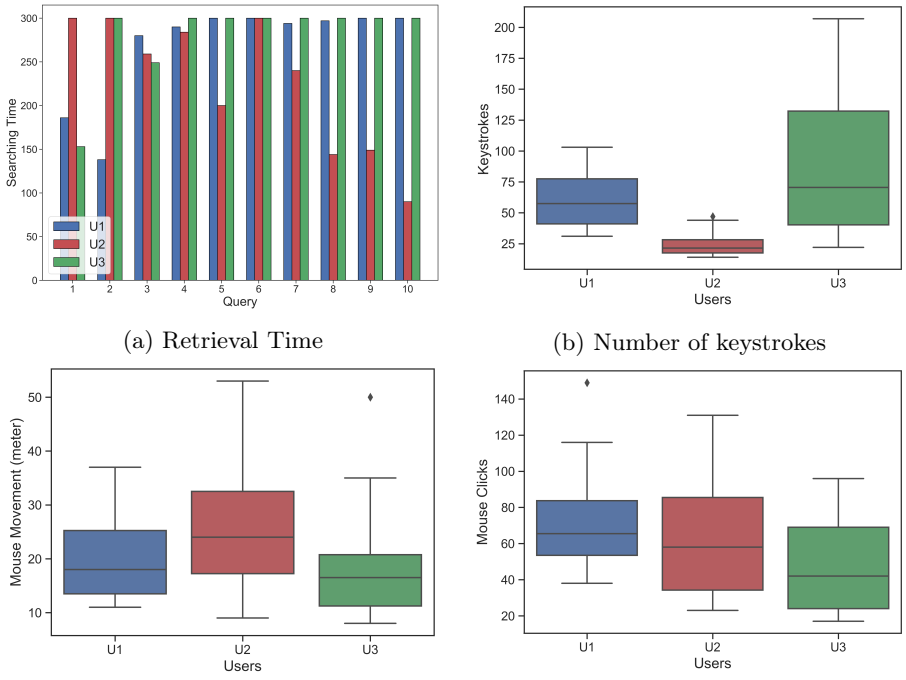
Fig. 4: Precision at 10 (P@10) and Cluster Recall at 10 (CR@10) of three users (U1: Developer, U2: Lifelogger, U3: Novice) for the original compared to the post-processing answers.

5.2 User Experiments

We asked the users to keep track of their searching time, the distance of the mouse movement (in meters), number of mouse clicks, and the number of keystrokes after they finished a question. We illustrate these results in Figure 5.

All participants tended to utilize the entire five minutes for doing a query as they are suggested. It is apparent that the lifelogger needed a shorter time to finish some topics than two others because this user already had unique information about the collection and hence could form better queries and knew when to stop searching. The mouse's distance moving on the screen, and the number of clicks reflected how users interacted with the user interface were also measured.

The former measurement of the author and the novice were similar while the data owner had a slightly higher number due to the larger widescreen monitor used in the experiment (note that screen-size has an effect). This statistic indicated the simplicity of our interface design. All features were shown clearly on the screen, and even those not familiar with the system could operate effectively. The lifelogger and the author clicked much more than the novice user. It may come from the fact that they both had sufficient experience to carefully check each image's full information prior to selecting it. In contrast, we observed that the novice user only checked the results on the main screen but not into the photo's details. The number of keystrokes of the developer was lower than that of the novice as the author knew what were the suitable keywords to find the answer, while the latter user had tried many concepts to be able to get the result. It was noticeable that the data owner had used a keyboard less frequently compared to other users when he did the experiment with the least number of keystrokes.



(a) Retrieval Time

(b) Number of keystrokes

(c) Mouse movement distance

(d) Number of clicks

Fig. 5: The evaluation of how three users (U1: Developer, U2: Lifelogger, U3: Novice) interacted with our system.

6 Discussion and Conclusions

The experiments showed that the expert user could achieve a similar result of the lifelogger who owns the dataset and also is the target of the system. Knowing the data well becomes the most significant advantage in the experiment. For instance, in the topic *"Find the moment when u1 was looking at lifelog data on a large touchscreen on the wall"*, the lifelogger did it instantly by using his prior knowledge to search for the particular location where this event took place. Meanwhile, both the creator of the system and the novice user did not have such insights. However, this merit was not enough to secure getting the high score as sometimes the data owner missed some relevant events in a topic. This issue is reasonable when a lifelogger usually has a massive dataset, and it is onerous to remember precisely without losing any moment within it. The system, in this scenario, could help its user to solve this problem. The first topic *"Find the moment when u1 was attending a praying rite with other people in the church"* witnessed the developer and the novice user gain a higher score than the lifelogger. The user experiment implied that there was almost no difference in the system manipulation between users. However, we have a noticeable gap in the scores of novice users compared to others. Another notable point is that the lifelogger and the novice users rarely used the "Word list" panel but tried to search for other concepts. This is perhaps an indication that the panel was not intuitive for non-expert users and that the users could have benefited from more training on the use of the system.

Considering opportunities for improvement, firstly the location information in the dataset seems to vary in accuracy, as stated by the lifelogger while testing. This issue became a critical problem when users wanted to retrieve within the specific area, such as a bus stop near their houses in topic 3 or churches in topic 1. Additionally, the detected concepts from Microsoft API service appeared with many too specific terms leading to the decrease in the precision while searching. There is a need for grouping these concepts to make the system more efficient. Another thing is that our object descriptor cannot recognize the colors well in the images, which is an essential feature in some cases like topic 10. In this work, we used this attribute from the Microsoft service, but it can be improvable by retraining our descriptor in some public datasets supporting these characteristics [10].

In this paper, we aimed to solve the challenge of retrieving an exact moment in a lifelogger lifetime from the large scale multimodal dataset. Our system is modified from the previous version, which combined the retrained images descriptor with the query expansion mechanism, by updating the additional visual similarity functions and reorganizing the user interface to be suitable for the ImageCLEFlifelog2020 LMRT challenge. The user experiment revealed that our simple designed system could be operated easily by the novice user. The system, being utilized in the best way, can help the developer obtain equivalent results with the lifelogger after our post-processing stage.

7 Acknowledgement

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. We acknowledge the support and input of the DCU ethics committee and the risk & compliance officer.

References

1. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
2. Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Zhou, L., Lux, M., Le, T.K., Ninh, V.T., Gurrin, C.: Overview of imagecleflifelog 2019: solve my life puzzle and lifelog moment retrieval. In: CLEF2019 Working Notes. CEUR Workshop Proceedings. vol. 2380, pp. 09–12 (2019)
3. Dao, M.S., Vo, A.K., Phan, T.D., Zettsu, K.: Bidal@ imagecleflifelog2019: the role of content and context of daily activities in insights from lifelogs. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org< http://ceur-ws.org (2019)
4. Doherty, A.R., Pauly-Takacs, K., Caprani, N., Gurrin, C., Moulin, C.J., O'Connor, N.E., Smeaton, A.F.: Experiences of aiding autobiographical memory using the sensecam. Human–Computer Interaction 27(1-2), 151–174 (2012)
5. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Albatal, R.: Ntcir lifelog: The first test collection for lifelog research. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 705–708 (2016)
6. Gurrin, C., Le, T.K., Ninh, V.T., Dang-Nguyen, D.T., Jónsson, B.Þ., Lokoš, J., Hürst, W., Tran, M.T., Schoeffmann, K.: Introduction to the third annual lifelog search challenge (lsc'20). In: Proceedings of the 2020 International Conference on Multimedia Retrieval. pp. 584–585 (2020)
7. Gurrin, C., Schoeffmann, K., Joho, H., Leibetseder, A., Zhou, L., Duane, A., Nguyen, D., Tien, D., Riegler, M., Piras, L., et al.: Comparing approaches to interactive lifelog search at the lifelog search challenge (lsc2018). ITE Transactions on Media Technology and Applications 7(2), 46–59 (2019)
8. Gurrin, C., Smeaton, A.F., Doherty, A.R., et al.: Lifelogging: Personal big data. Foundations and Trends® in information retrieval 8(1), 1–125 (2014)
9. Ionescu, B., Müller, H., Péteri, R., Abacha, A.B., Datla, V., Hasan, S.A., Demner-Fushman, D., Kozlovski, S., Liauchuk, V., Cid, Y.D., Kovalev, V., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G.: Overview of the ImageCLEF 2020: Multimedia retrieval in lifelogging, medical, nature, and internet applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), vol. 12260. LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece (September 22-25 2020)

10. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision 123(1), 32–73 (2017)
11. Le, N.K., Nguyen, D.H., Nguyen, V.T., Tran, M.T.: Lifelog moment retrieval with advanced semantic extraction and flexible moment visualization for exploration. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEURWS. org< http://ceur-ws. org (2019)
12. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. vol. 2, pp. 1150–1157 vol.2 (1999)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (Nov 2004)
14. Luo, H., Wei, H., Lai, L.L.: Creating efficient visual codebook ensembles for object categorization. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 41(2), 238–253 (2011)
15. Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T.: Overview of ImageCLEF Lifelog 2020:Lifelog Moment Retrieval and Sport Performance Lifelog. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Thessaloniki, Greece (September 22-25 2020)
16. de Oliveira Barra, G., Cartas Ayala, A., Bolaños, M., Dimiccoli, M., Giró Nieto, X., Radeva, P.: Lemore: A lifelog engine for moments retrieval at the ntcir-lifelog lsat task. In: Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies (2016)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Tran, L.D., Nguyen, M.D., Binh, N.T., Lee, H., Gurrin, C.: Myscéal: An experimental interactive lifelog retrieval system for lsc'20. In: Proceedings of the Third Annual Workshop on Lifelog Search Challenge. pp. 23–28 (2020)
19. Zhou, L., Dang-Nguyen, D.T., Gurrin, C.: A baseline search engine for personal life archives. In: Proceedings of the 2nd Workshop on Lifelogging Tools and Applications. pp. 21–24 (2017)
20. Zhou, P., Bai, C., Xia, J.: Zjutcvr team at imagecleflifelog2019 lifelog moment retrieval task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org< http://ceur-ws. org (2019)