LifeSeeker 2.0 : Interactive Lifelog Search Engine at LSC 2020

Tu-Khiem Le[*]
Van-Tu Ninh[*]
Dublin City University
Ireland

Minh-Triet Tran
Thanh-An Nguyen
Hai-Dang Nguyen
University of Science
Vietnam National University Ho Chi Minh city
Vietnam

Liting Zhou
Graham Healy
Cathal Gurrin
Dublin City University
Ireland

ABSTRACT

In this paper we present our interactive lifelog retrieval engine in the LSC'20 comparative benchmarking challenge. The LifeSeeker 2.0 interactive lifelog retrieval engine is developed by both Dublin City University and Ho Chi Minh University of Science, which represents an enhanced version of the two corresponding interactive lifelog retrieval engines in LSC'19. The implementation of LifeSeeker 2.0 has been designed to focus on the searching by text query using a Bag-of-Words model with visual concept augmentation and additional improvements in query processing time, enhanced result display and browsing support, and interacting with visual graphs for both query and filter purposes.

CCS CONCEPTS

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

KEYWORDS

lifelog, interactive retrieval, information system

ACM Reference Format:
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 the Lifelog Search Challenge (LSC '20), June 9, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3379172.3391724

1 INTRODUCTION

Information retrieval systems can be effective tools to support our daily search activities and are gradually becoming an essential part of modern life. Some typical examples are search engines and product search systems in online stores. The information retrieval system can range from straightforward retrieval of information from structured databases based on some predefined filtering conditions, to a complicated search system which requires multiple types of query inputs (text description, gps, heart rate, location categories, etc.) to retrieve a desired result. Despite the complexity of any information retrieval system, it often comprises two components: the indexing system and the query system [22]. The challenge of the indexing system is to analyze and extract important features that the query system can use to retrieve the targeted information effectively.

For the query system, most are designed for interactive usage which requires the conversion from personal self-analysis and context understanding of the query into a proper input to the system. As the user is a key component of the system, many challenges are posed to both design an effective interactive user interface and query methods aiming to support the novice user, and create an indexed database to deal with complex queries which require multi-modal conditions for retrieval.

Since the seminal MyLifeBits [8] lifelog database in 2006, lifelogging has gradually become an active research topic. Many active challenges and tasks have been proposed to explore ways of deriving insights of an individual's life by using lifelog data. Most of the challenges focus on the task of lifelog moment retrieval which aims to build lifelog retrieval systems to find specific moments in an individual's life as well as providing an understanding of individual's habits and activities [4–6, 10–13]. The Lifelog Search Challenge (LSC) is an international competitive benchmarking activity with the aim of supporting the fair and accurate comparison of different approaches to interactive retrieval from lifelog datasets [12].

In this paper, we describe our lifelog interactive retrieval system - LifeSeeker 2.0 - participating in LSC'20. We provide detailed information of retrieval methods, user interface, lifelog data analysis and indexed database design used. The LifeSeeker 2.0 inherits many features from prior research of interactive lifelog retrieval systems but improves upon these by using additional novel features such as augmented visual concepts extracted from images to enrich the data descriptions, elastic sequencing to use temporal information of nearby moments, and a three-granularity-level graph-based location filtering interface. These features were developed to facilitate the use of both expert and novice users to increase the efficiency of user interaction with the multi-modal lifelog retrieval system. We describe the essential novel features and show how LifeSeeker 2.0 operates in this paper.

[*] Both authors contributed equally to this research.

LSC '20, June 9, 2020, Dublin, Ireland
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7136-0/20/06…$15.00
https://doi.org/10.1145/3379172.3391724

2 RELATED RESEARCH

In the recent years, there has been much research focused on developing a better understanding of ways to interact with lifelog data, and subsequently, there has a large amount of research focused on developing information retrieval approaches to recall specific moments within a provided set of lifelog data. In order to compare the performance of retrieval systems, various tasks have been organised such as the NTCIR14-Lifelog task [11], ImageCLEF lifelog task [5], and the Lifelog Search Challenge (LSC) [13], where each evaluates the systems using different metrics. Although LSC is a new challenge that occurred only in 2018 and 2019, it has been gaining more attention as it was specifically designed to compare approaches in real-time. Nine teams participated last year in LSC (2019) and we would like to highlight three systems that achieved the best results.

The first system is vitrivr [24] - a multimedia retrieval system that is built to cope with many types of media including images, videos, audios and 3D models. It was the winner of the Video Browser Showdown [25] in 2017 and 2019, which is a similar challenge to LSC, but is aimed at interactive video retrieval. The LSC2019 vitrivr system introduced a new type of media data called image sequence that combines a series of images into segments, which enables a better data representation and association. The engine was also equipped boolean querying as a late-filtering approach to fuse the results. VIRET [19] is another video retrieval system at VBS which proved to work perfectly on lifelog data by considering day and image of a day as "video" and "shot" correspondingly. In LSC2019, VIRET was upgraded with a query panel to allow better filtering by the metadata (week days, heart rate, time) and the GPS locations to help the system narrow down the images to search for. The HCMUS team [17] attempted to increase the search accuracy by enhancing the number of concepts using multiple detectors. The authors developed a special concept detector based on the habits of the lifelogger on daily basis, this resulted in rich and accurate metadata. The data is hashed into a table and then converted into a tree structure, that can be retrieved by the search engine to support synonym search and autocomplete component.

3 LIFELOG DATA FOR THE EXPERIMENT

The LSC'20 dataset is a new multimodal dataset that combines the datasets from three lifelog moment retrieval tasks part of prior NTCIR challenges from 2015, 2016, and 2018 [12]. This dataset is a 114-day mutimodal lifelog dataset of one individual who wore multiple sensors and used a smartphone to capture the data continuously. The images of the lifelog data are redacted to blur faces and remove essential individual's textual content to protect the privacy of the lifelogger and others. The metadata of the data is enriched and refined to provide as much information as possible for data processing and analysis to support multiple interactive retrieval system designs. Moreover, the data is enhanced with the automatic extraction of location attributes and location categories, and visual objects using the output of computer-vision neural networks such as Place365CNN [27] and Mask-RCNN [14, 26] that has pre-trained on the COCO dataset using 80 items [18]. This data is available for participants to download[1] and is used for the live search challenge with newly generated topics for both expert and novice users.

4 OVERVIEW OF LIFESEEKER 2.0

Since the LifeSeeker 2.0 interactive retrieval system inherits many features from the first version used in LSC'19, we briefly describe some essential features of the LifeSeeker interactive retrieval system from LSC'19 in Section 4.1 and list the features that we reuse and upgrade in LifeSeeker 2.0 from the its first version. Then, we present the novel features with detailed enhancement of LifeSeeker 2.0 in LSC'20 compared to LifeSeeker in LSC'19.

Figure 1 illustrates the main interface of both free text search and its results. The results are a ranked list of moments (a ranked list of single images), which is relevant to the query, shown in a grid list. The expandable box with four sections of images below the grid list show more related moments which are relevant to the chosen image. They include: left - the moments before the chosen image; right - the moments after the chosen image; top - the pinned images to save potential relevant moments; and bottom - a ranked list of images which are visually similar to the current one. Figure 3 demonstrates another interface which integrates clustered location-based search and filtering. The left collapsible window shows the movement timeline of the lifelogger displayed as a graph. The right menu shows images relevant to corresponding locations. Users can interact with the graph to search or filter to retrieve the results.

In summary of the features used in LifeSeeker 2.0 , it includes two query modes which are search mode (search using free-text and location conditions) and filter mode (filter using time and location keywords in free-text search box or use graph-based filter). To verify a certain moment, elastic sequencing is attached to the system to view the before and after events of a that moment based on a defined time difference from the current moment, for instance, five minutes before and after the chosen moment. A moment is a single image with multimodal metadata showing relevant information of actions, locations and other related information at a point of time. It is different from an event, which is a sequence of moments showing relevant actions in a certain location during a range of time.

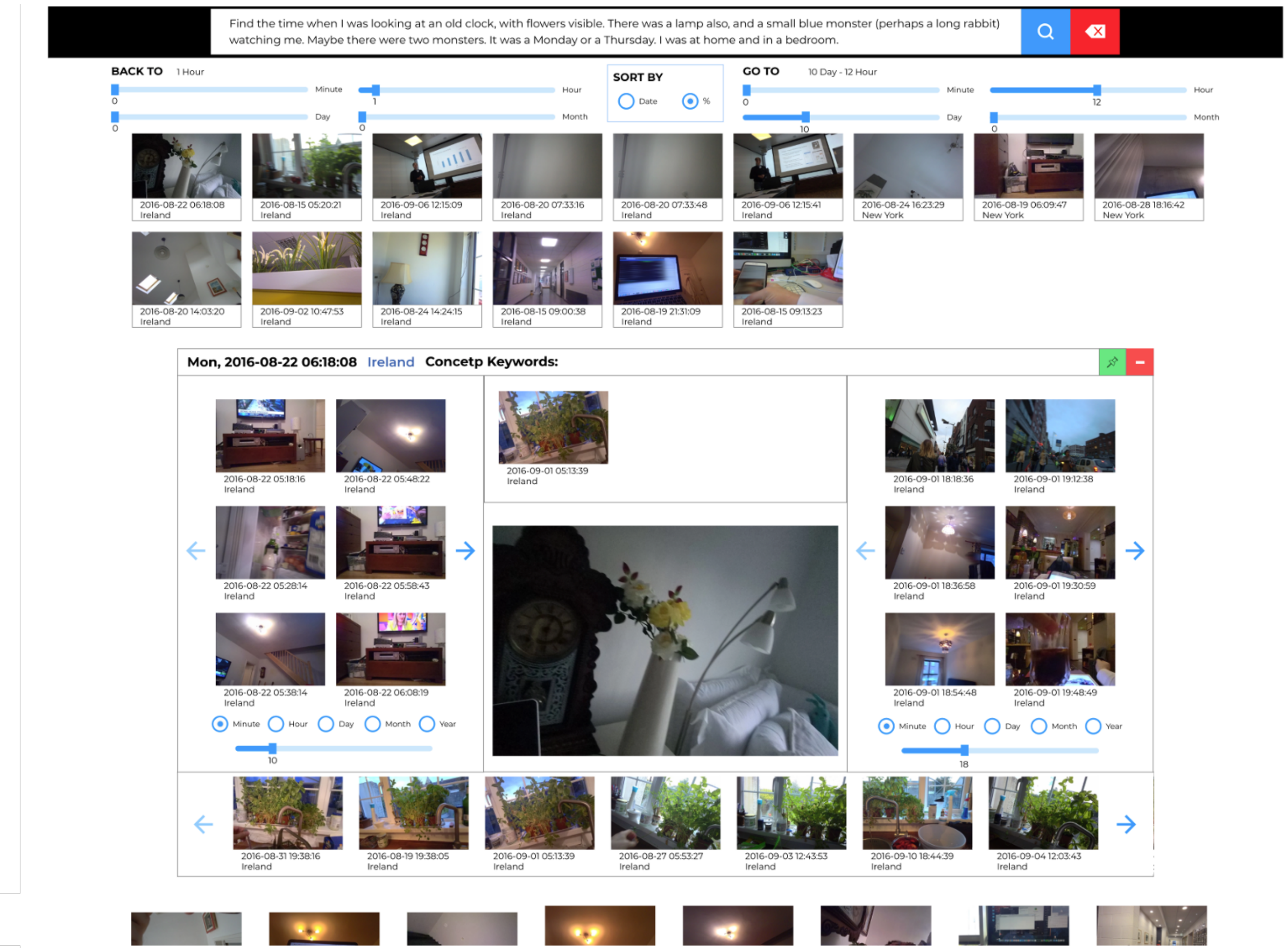
We also enhance the metadata to provide more concepts from the images by using many different detectors, which will be described in the following section. All enhanced data employed is indexed before querying or ranking the results.

4.1 Overview of LifeSeeker in LSC'19

The LifeSeeker interactive retrieval engine is an enhancement from the baseline version used in the NTCIR-14 Lifelog3 task, incorporating careful qualitative user study feedback from four novice users [21].

The user interface was designed to be used by novice users and thus incorporates a simple faceted filtering mechanisms that allows users to find moments using predefined conditions for date, time, location, visual concepts, heart rate, etc. and a simple text search box to input text queries. Retrieval results are presented as a ranked list of relevant images shown in card views. The user can browse, find content-similarity images from a specific image, and narrow the search results with the faceted filter panel.

[1] LSC Website: http://lsc.dcu.ie



Find the time when I was looking at an old clock, with flowers visible. There was a lamp also, and a small blue monster (perhaps a long rabbit) watching me. Maybe there were two monsters. It was a Monday or a Thursday. I was at home and in a bedroom.

BACK TO 1 Hour

SORT BY ○ Date ● %

GO TO 10 Day - 12 Hour

Mon, 2016-08-22 06:18:08 Ireland Concetp Keywords:

Minute
Hour
Day
Month
0
1
10
12
2016-08-22 06:18:08 Ireland
2016-08-15 05:20:21 Ireland
2016-09-06 12:15:09 Ireland
2016-08-20 07:33:16 Ireland
2016-08-20 07:33:48 Ireland
2016-09-06 12:15:41 Ireland
2016-08-24 16:23:29 New York
2016-08-19 06:09:47 New York
2016-08-28 18:16:42 New York
2016-08-20 14:03:20 Ireland
2016-09-02 10:47:53 Ireland
2016-08-24 14:24:15 Ireland
2016-08-15 09:00:38 Ireland
2016-08-19 21:31:09 Ireland
2016-08-15 09:13:23 Ireland
2016-08-22 05:18:16 Ireland
2016-08-22 05:48:22 Ireland
2016-08-22 05:28:14 Ireland
2016-08-22 05:58:43 Ireland
2016-08-22 05:38:14 Ireland
2016-08-22 06:08:19 Ireland
● Minute ○ Hour ○ Day ○ Month ○ Year
10
2016-09-01 05:13:39 Ireland
2016-09-01 18:18:36 Ireland
2016-09-01 19:12:38 Ireland
2016-09-01 18:36:58 Ireland
2016-09-01 19:30:59 Ireland
2016-09-01 18:54:48 Ireland
2016-09-01 19:48:49 Ireland
● Minute ○ Hour ○ Day ○ Month ○ Year
18
2016-08-31 19:38:16 Ireland
2016-08-19 19:38:05 Ireland
2016-09-01 05:13:39 Ireland
2016-08-27 05:53:27 Ireland
2016-09-03 12:43:53 Ireland
2016-09-10 18:44:39 Ireland
2016-09-04 12:03:43 Ireland

Figure 1: LifeSeeker 2.0 at LSC'20 - Main Interface

For retrieval, we employ a Bag-of-Words model for text retrieval. The dictionary of the model was created by combining text data from all metadata sources (visual concepts, location name, timezone). To reduce the lexical gap, we performed query expansion on the original dictionary using a Word2Vec model [20] pre-trained on GoogleNews dataset. Both text description of the images and query were encoded into a Bag-of-Words vector space for similarity comparison. The user can also query as well as filtering the results by using faceted filtering panel.

In the user interface of the LifeSeeker 2.0 interactive retrieval engine, we keep the search text box but enhance this with a term suggestion feature. For the retrieval mechanism, we implement a filter mechanism before using a Bag-of-Words model to rank and choose relevant moments. We extract the noun phrases and compare with three refined auto-generated dictionaries (time, location, and concepts) in descending priority order to filter the search results before encoding the relevant moments into the Bag-of-Words vector for re-ranking and generating the final top k results.

4.2 Metadata Enhancement

We observe that most of the valuable information in lifelog data is obtained using visual features that are extracted from the images. The more features we can extract from these lifelog images, the better the match between a user query and the indexed terms. In the provided metadata within the dataset, the organisers include place attributes and visual objects, and these are extremely useful sources of evidence. However, the number of object categories in this metadata is small with only 80 classes (from a network trained using the COCO dataset [18]) whereas the number of diverse objects appearing in lifelog data is typically much larger. Therefore, in order to capture more objects, we utilise the bottom up attention model [1] which is based on Faster R-CNN [23] with ResNet101 [15] and pre-trained on the Visual Genome dataset [16]. This model is highly beneficial since it not only contains 1,600 object classes, but also describes the associating object attribute (with 400 attribute types). Figure 2 shows some example results of the object detection based on the Visual Genome dataset.

We also noticed that besides the visual data, the textual information in the images is of importance. The text from images can tell

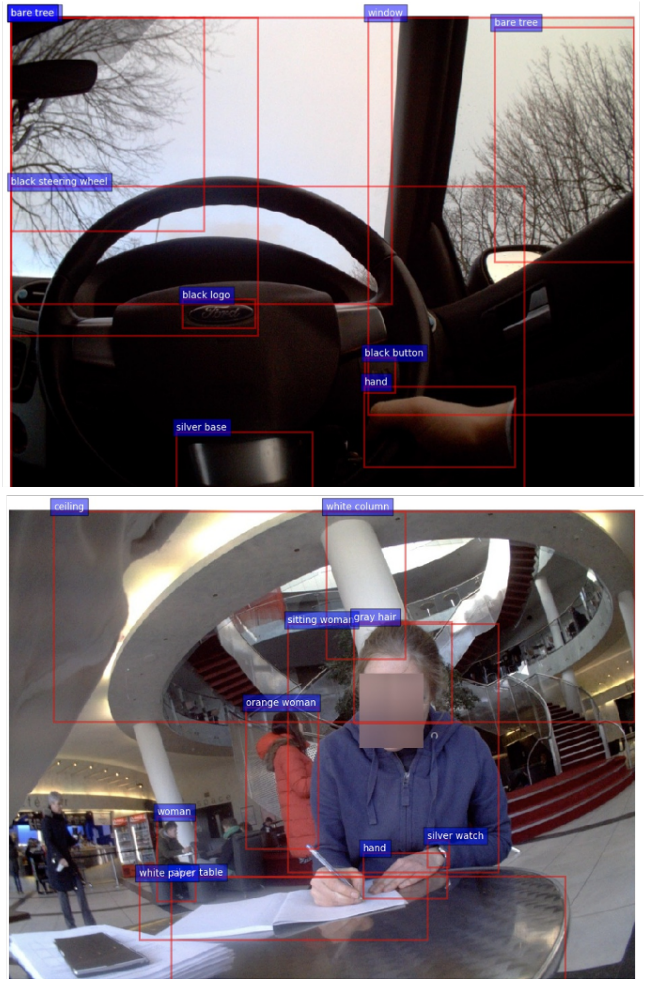


Figure 2: Visual Genome's objects in lifelog dataset. For example, we can see that the *black logo, silver base* and *bare tree* has been detected from the upper image and the *silver watch, white column* and *orange woman* in the lower image.

what the lifelogger is reading on a computer screen, what the brand name of the coffee shop is, etc. which is a great clue to identify the activities of lifelogger. To obtain text features, we utilised the work from [2] for text recognition and CRAFT [3] for scene text detection.

By aggregating the extra object categories (with attribute) with text extracted from scene and original data, we are able to craft a richer and more precise metadata for indexing the LSC dataset.

4.3 Additional Free-Text Ranking Modes

Apart from our free-text ranking method (BOW) implemented in previous version, in LifeSeeker 2.0 we introduce two more methods for indexing and searching for lifelog moments which are the Elastic Search (ES) and Visual Vector (VV).

In the ES mode, we utilise the Elastic Search [9] engine which is an open-source software for indexing, analyzing and retrieving data. The aim of Elastic Search is speed and scale which aligns with our purpose - to boost the overall retrieval speed of LifeSeeker. The database was created based on our enhanced metadata. A query in ES mode will be pre-processed the same way, with BoW, which means that the stop-words will be removed and the while the remaining words will go through the stemming process to obtain their stemmed form. After that, we used Query DSL (Domain Specific Language) provided by Elastic Search to generate queries for the search engine.

The VV mode offers a different approach to retrieve potentially relevant content by comparing the distance between visual vectors of the query and the images. Firstly, ResNet101 features for all images in the dataset were extracted. Then W2VV [7] was used to convert a text query into visual vector space and produce a feature vector similar to that of ResNet101. The relevant images could be obtained by comparing the cosine similarity between these vectors.

4.4 Elastic Sequencing

It is our conjecture that target memory is usually retrieved by connecting the previous memories, which forms a path which leads to the piece of memory we want to recall. Taking an example query from last year's LSC: "Checking out of a hotel in the early morning, before six am.", the action "check out" is not visible nor identifiable by looking at the images separately. However, we could recognise the action by looking further in the past and checking if the lifelogger came the airport, got on a plane and arrived at a hotel, only then can we tell that the future images will contain a checking-out activity.

Thus, in order to confirm the correctness of the image returned, we introduced an upgrade to our LifeSeeker engine to show a details view which displays next and previous images with respect the current image in a sequence. A scaling factor is also applied to control how far in time we would like to explore these images. By adjusting the time delta, images before and after the target can be adjusted to be temporally nearby or further apart.

4.5 Location-based Clustering

To support interactive graph-based search and filtering based on location, we propose to cluster the location names extracted from the metadata as well as GPS into three granularity levels: country, location and area. As GPS can only provide us with the location information without distinguishing between different areas in the same location, we wanted to create a finer level of detail about location information. Therefore, we manually labelled the type of area in the location for some images, then assigned the remaining unlabelled images into the predefined areas by considering visual similarity score, the number of similar location attributes and location categories of the unlabelled images compared to the ones of each area. Some of the object concepts which are visually detected by Mask-RCNN [14] are also considered after they are refined to support the type of areas to which images belong. For example, GPS only provides us the information that the lifelogger is at home, however, it does not provide any detail as to whether the lifelogger is currently in the living room, the kitchen, the bathroom or the bedroom. Some typical objects associated with specific areas in a location, such as the presence of an oven and stove in the

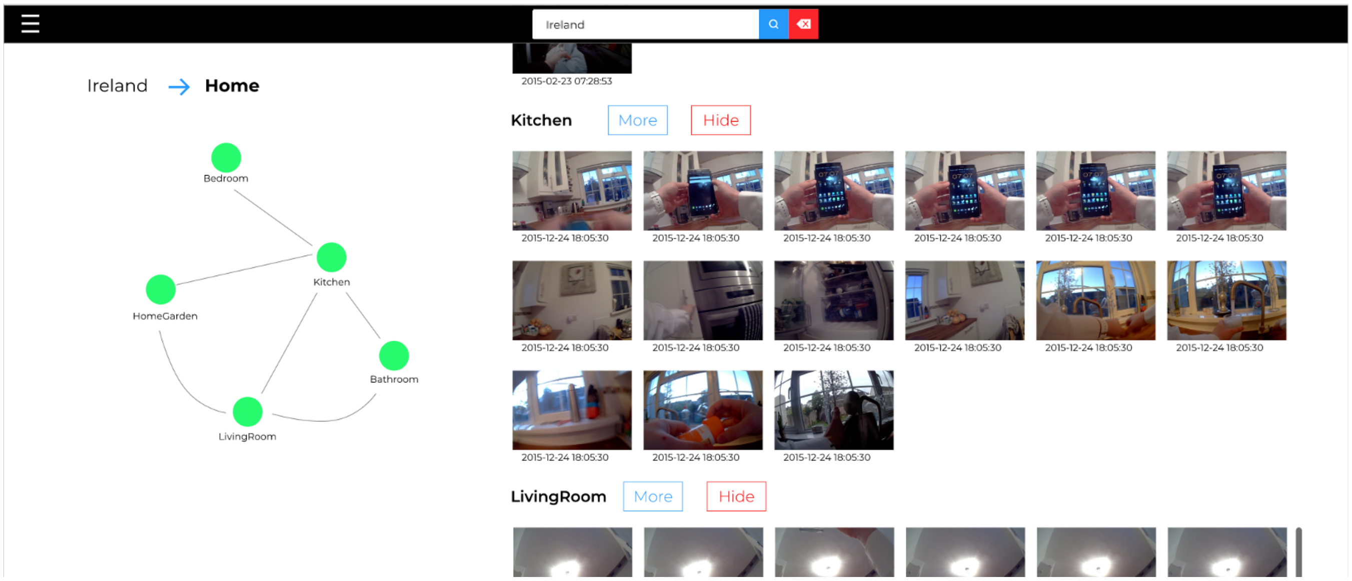


Figure 3: LifeSeeker 2.0 at LSC'20 - Location Search and Filter Interface

kitchen; a sofa and a television in the living room or a mirror above a sink in the bathroom, are useful to infer the type of areas in that location. From these three granularity levels, a user can interact with a graph-based visualization from the country (first level) to the location (second level) and finally to the area of the location if necessary (third level) to filter and query. In this way, a user can have an overall summary of the places and interact quickly to filter the results.

4.6 Interactive Transitional Graph-based Filter

To enhance the visualization of the graph-based filter after clustering the location into three granularity levels, we add the chronological order of places visited to the graph. This means that we add directional edges between the locations in the graphs. As it can be a fully connected graph, a user can choose the location/area first, then only the links to the chosen location/area are kept and highlighted with the corresponding connected nodes. Since most of the queries describe the chronological order of the actions which involve many typical corresponding locations, we exploit this feature so that a user can verify the filtering option quickly. It is important because we do not want users, especially novice ones, to waste time on irrelevant filtering options which have the same location, area and time but have a different date or have chronological activities in different areas.

Figure 4 illustrates a prototype of this feature in our system. The green node is the current chosen location. The related edges are highlighted in red so that the user can focus on the history of visited places in chronological order and continue to select the next location to narrow down the scope of the search by the temporal order of locations. The grey edges are not related to the chosen green node. Some of the grey edges might turn to red if they connect to the next chosen node.

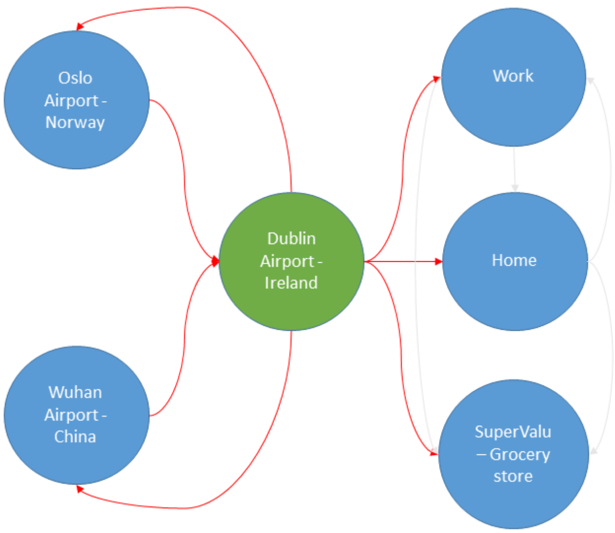


Figure 4: Illustration of the interactive graph-based filtering with transitional edges displaying chronological order of visited places.

4.7 Common Movement Patterns

A lifelogger usually has common working/movement patterns that can be detected to support further query. By utilising the location-based clustering and temporal information of images, our system can detect common movement patterns, such as going from home to work, or from a lecture room to a favourite restaurant, etc. Our system can suggest to a user some common movement patterns to assist the user to handle movement-related queries.

5 CONCLUSION

In this paper, we presented an overview of the changes made to our interactive lifelog retrieval engine LifeSeeker 2.0 . LifeSeeker 2.0 was enhanced in not only the user interface, but also the search engine. We note that LifeSeeker 2.0 is able to integrate more concepts from the lifelog data and able to deliver a better search results.

For future developments of the system, we are aiming to create a self-learning mechanism so that the system could learn from labels requested for unknown objects. We expect this will boost the performance of the engine as it would enable us to capture keywords in the query better and mitigate noise introduced from query expansion. Moreover, we also planning to conduct a number of experiments on the user interface by getting participants perform different types of searches. Observations on the habits of how a user conducts searches on our engine will give us information on where to optimise the interface so that the search process will be quicker and more accurate.

ACKNOWLEDGMENTS

This publication has emanated from research supported in party by research grants from Irish Research Council (IRC) under Grant Number GOIPG/2016/741, Science Foundation Ireland under grant numbers SFI/12/RC/2289 and SFI/13/RC/2106, and Vietnam Ireland Bilateral Education Exchange Programme (VIBE) 2019. We would like to acknowledge Minh-Huy Nguyen-Ngoc, Huu-Duc Trinh and Nguyen-Hien Tran for supporting us with the interface of the system.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*.

[2] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. 2019. What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. In *International Conference on Computer Vision (ICCV)*. to appear.

[3] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character Region Awareness for Text Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9365–9374.

[4] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Giulia Boato, Liting Zhou, and Cathal Gurrin. 2017. Overview of ImageCLEFlifelog 2017: Lifelog Retrieval and Summarization. In *CLEF2017 Working Notes (CEUR Workshop Proceedings)*. CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland.

[5] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Minh-Triet Tran, Liting Zhou, Mathias Lux, Tu-Khiem Le, Van-Tu Ninh, and Cathal Gurrin. 2019. Overview of ImageCLEFlifelog 2019: Solve my life puzzle and Lifelog Moment Retrieval. In *CLEF2019 Working Notes (CEUR Workshop Proceedings)*. CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland.

[6] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. 2018. Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval. In *CLEF2018 Working Notes (CEUR Workshop Proceedings)*. CEUR-WS.org <http://ceur-ws.org>, Avignon, France.

[7] Jianfeng Dong, Xirong Li, and Cees G. M. Snoek. 2018. Predicting Visual Features From Text for Image and Video Caption Retrieval. *IEEE Transactions on Multimedia* 20, 12 (2018), 3377–3388.

[8] Roger Gemmell, Jim; Bell, Gordon; Lueder. 2006. My lifebits: a personal database for everything. *Commun. ACM* 49, 1 (2006), 88–95. https://doi.org/10.1145/1107458.1107460

[9] Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide* (1st ed.). O'Reilly Media, Inc.

[10] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, and Rami Albatal. 2016. Overview of NTCIR-12 Lifelog Task. (2016), 354–360. http://eprints.gla.ac.uk/131460/ The authors acknowledge the financial support of Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289 and the input of the DCU ethics committee and the risk &amp; compliance officer. We acknowledge financial support by the European Science Foundation via its Research Network Programme ?Evaluating Information Access Systems?.

[11] Cathal Gurrin, H. Joho, Frank Hopfgartner, Liting Zhou, Tu Ninh, Tu-Khiem Le, Rami Albatal, D.-T Dang-Nguyen, and Graham Healy. 2019. Overview of the NTCIR-14 Lifelog-3 task.

[12] 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.

[13] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Bernd Munzer, Rami Albatal, Frank Hopfgartner, Liting Zhou, and Duc-Tien Dang-Nguyen. 2019. A Test Collection for Interactive Lifelog Retrieval. In *MultiMedia Modeling*, Ioannis Kompatsiaris, Benoit Huet, Vasileios Mezaris, Cathal Gurrin, Wen-Huang Cheng, and Stefanos Vrochidis (Eds.). Springer International Publishing, 312–324.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR* abs/1703.06870 (2017). arXiv:1703.06870 http://arxiv.org/abs/1703.06870

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. https://arxiv.org/abs/1602.07332

[17] Nguyen-Khang Le, Dieu-Hien Nguyen, Trung-Hieu Hoang, Thanh-An Nguyen, Thanh-Dat Truong, Duy-Tung Dinh, Quoc-An Luong, Viet-Khoa Vo-Ho, Vinh-Tiep Nguyen, and Minh-Triet Tran. 2019. Smart Lifelog Retrieval System with Habit-Based Concepts and Moment Visualization. In *Proceedings of the ACM Workshop on Lifelog Search Challenge* (Ottawa ON, Canada) *(LSC '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3326460.3329155

[18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014). arXiv:1405.0312 http://arxiv.org/abs/1405.0312

[19] Jakub Lokoč, Tomáš Souček, Premysl Čech, and Gregor Kovalčík. 2019. Enhanced VIRET Tool for Lifelog Data. In *Proceedings of the ACM Workshop on Lifelog Search Challenge* (Ottawa ON, Canada) *(LSC '19)*. Association for Computing Machinery, New York, NY, USA, 25–26. https://doi.org/10.1145/3326460.3329159

[20] Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. 1–12.

[21] Van-Tu Ninh, Tu-Khiem Le, Liting Zhou, Graham Healy, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Sinead Smyth, and Cathal Gurrin. 2019. A Baseline Interactive Retrieval Engine for the NTICR-14 Lifelog-3 Semantic Access Task. In *The Fourteenth NTCIR conference (NTCIR-14)* (Tokyo, Japan).

[22] Mario Pérez-Montoro and Lluís Codina. 2017. Chapter 5 - The Essentials of Search Engine Optimization. In *Navigation Design and SEO for Content-Intensive Websites*, Mario Pérez-Montoro and Lluís Codina (Eds.). Chandos Publishing, 109 – 124. https://doi.org/10.1016/B978-0-08-100676-4.00005-5

[23] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *CoRR* abs/1506.01497 (2015). arXiv:1506.01497 http://arxiv.org/abs/1506.01497

[24] 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* (Ottawa ON, Canada) *(LSC '19)*. Association for Computing Machinery, New York, NY, USA, 27–31. https://doi.org/10.1145/3326460.3329160

[25] Klaus Schoeffmann. 2019. Video Browser Showdown 2012-2019: A Review. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. 1–4. https://doi.org/10.1109/CBMI.2019.8877397

[26] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

[27] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).