Flexible Interactive Retrieval SysTem 2.0 for Visual Lifelog Exploration at LSC 2021

Hoang-Phuc Trang-Trung[1,2,3], Thanh-Cong Le[1,2,3], Mai-Khiem Tran[1,2,3], Van-Tu Ninh[4], Tu-Khiem Le[4], Cathal Gurrin[4], Minh-Triet Tran[1,2,3*]
[1]University of Science, VNU-HCM, Vietnam
[2]John von Neumann Institute, VNU-HCM, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
[4]Dublin City University, Ireland

ABSTRACT

With a huge collection of photos and video clips, it is essential to provide an efficient and easy-to-use system for users to retrieve moments of interest with a wide variation of query types. This motivates us to develop and upgrade our flexible interactive retrieval system for visual lifelog exploration. In this paper, we briefly introduce version 2 of our system with the following main features. Our system supports multiple modalities for interaction and query processing, including visual query by meta-data, text query and visual information matching based on a joint embedding model, scene clustering based on visual and location information, flexible temporal event navigation, and query expansion with visual examples. With the flexibility in system architecture, we expect our system can easily integrate new modules to enhance its functionalities.

CCS CONCEPTS

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

KEYWORDS

lifelog, interactive retrieval, information system, joint embedding model, component integration

ACM Reference Format:
Hoang-Phuc Trang-Trung, Thanh-Cong Le, Mai-Khiem Tran, Van-Tu Ninh, Tu-Khiem Le, Cathal Gurrin, Minh-Triet Tran. 2021. Flexible Interactive Retrieval SysTem 2.0 for Visual Lifelog Exploration atLSC 2021. In *Proceedings of the 4th Annual Lifelog Search Challenge (LSC'21), August 21, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 6 pages.
https://doi.org/10.1145/3463948.3469072

1 INTRODUCTION

People create a huge collection of photos and videos to capture daily activities as well as special moments in their lives. Such visual and related metadata are valuable sources for lifelog retrieval[10], not only to revive memories or to verify events but also to analyze for a better understanding of our behaviors and habits[11]. This can help re-design business processes, create better personalized intelligent services, improve personal lifestyle and health, etc.

Lifelog data consists of data in different formats, such as photos, video clips, recorded audio clips, GPS information, personal healthcare data, data from different sensors, etc. To efficiently retrieve and analyze lifelog data, it is necessary to devise and develop a flexible interactive retrieval system that can support users to input their queries in different modalities. Therefore, the annual Lifelog Search Challenge (LSC[10]) has been organized to encourage different research groups worldwide to enhance their solutions for interactive lifelog retrieval systems.

Because of the wide variation in query types, we build a flexible query system that can integrate new query processing components to handle various query types. Our platform can also integrate different AI services to analyze lifelog data, *e.g.*object detection, scene classification and attribute detection, image captioning, *etc.* We also propose a mechanism to define service integration processes to deal with new scenarios for data analysis and query processing. Our Flexible Interactive Retrieval SysTem (FIRST version 2.0) provides the following main features for lifelog moment retrieval.

- Query by meta-data: Our system allows users to query with meta-data, including date, time, and location. We also extract scene text, entities, activities, places to enrich meta-data for each photo/scene.
- Text query and visual information matching based on joint embedding model: Our system encodes both a text query and a photo into an embedding space to measure their similarity.
- Scene clustering based on visual and location information: The system visualizes scene clusters with two main clustering conditions: visual similarity and (GPS) location.
- Flexible temporal event navigation: Users can navigate visual lifelog photos with a flexible timeline to shrink or expand the time interval of interest.
- Query expansion with visual examples: From one or several example photos, our system can retrieve and rank photos based on their similarity and dissimilarity.

The content of this paper is organised as follows. In Section 2, we briefly review related approaches for lifelogging retrieval. We introduce an overview and main components of our system in Section 3. We then illustrate some usage scenarios of our system in Section 4. The conclusion and discussion for future work are in Section 5.

*Corresponding author: tmtriet@fit.hcmus.edu.vn

LSC '21, August 21, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8533-6/21/08…$15.00
https://doi.org/10.1145/3463948.3469072

2 RELATED WORK

Lifelog retrieval requires visual data understanding and an efficient method for users to interact with a query system. For ImageCLEF lifelog tasks [6, 25] or NTCIR14-Lifelog task [9], the objective is to evaluate the retrieval methods to find accurate results corresponding to a query. For Lifelog Search Challenge (LSC[1, 2, 10]), the goals focus on both lifelog data understanding and efficiency of user interaction with retrieval systems. This year, the fourth Lifelog Search Challenge [12] will also take place in ICMR'21 conference.

To effectively index the lifelog images, many systems utilize the existing pretrained models on various computer vision tasks like object detection, scene classification and scene text detection to detect popular concepts in images. Then they use those indices to retrieve the correct moments by analyzing the query in an interactive manner. Throughout the years, many improvements have been brought out to make lifelog systems more accurate and easy-to-use.

The vitrivr system [13] is a multimedia retrieval system that supports different media types. The system for LSC 2020 integrates two new features, temporal scoring and staged querying. Another video search system, VIRET [15], is used in this lifelog search problem with advanced tools to browse images in temporal domain.

In Mysceal [29], the authors introduce a TF-IDF to match free-text information query with lifelog index.

The LifeGraph [27] authors build a knowledge graph linking lifelog data to external data to capture broader context.

The interesting function of Exquisitor [14] is the capability to use relevance feedback, including positive and negative examples, to learn the needs from users and to refine the results. We inherit this function in our system for query expansion.

In SOM Hunter system [24], the semantic distance between a query, in text format, and a photo/video clip is measured in a common space. Therefore, this retrieval tool is a successful system for known-item and ad-hoc-search tasks. Both LifeSeeker 2.0 [18] and FIRST v1 [30] also support this approach. LifeSeeker 2.0 focuses on text query using a Bag-of-Words approach with visual concept augmentation[18], while our previous systems [16, 30] exploit personalized concepts.

Besides traditional ways to input query information, new methods for interaction have been developed to assist users, such as voice-controlled interactive search engine in Voxento [3] or natural interaction in virtual reality environment of VRLE system [7].

Our text-based image retrieval feature is based on the image-text matching methods, which can be divided into two categories: (1) building a network to predict the similarity score between image and text pairs [19, 21] or (2) projecting the image and text embeddings onto a joint latent space under which we can directly compare image and text representations [5, 8, 20].

To prevent running a big model for all the images at test time, we employ the latter approach to design the architecture for our image-text matching model. In recent years, this field has further improved with image-text pretraining paradigm [4, 21]. Those models are trained on large datasets using contrastive loss to understand the image and text representations better. Then they are applied to downstream tasks like image captioning, visual question answering, or image-text retrieval and achieve great results.

3 LIFELOG EVENT RETRIEVAL WITH FLEXIBLE COMPONENTS

3.1 System Overview

Figure 1 shows the overview structure of our proposed system. We use the user-interface and system integration platforms from our FIRST version 1 [30] to manage different layouts and user interaction modalities, and query processing components, respectively.

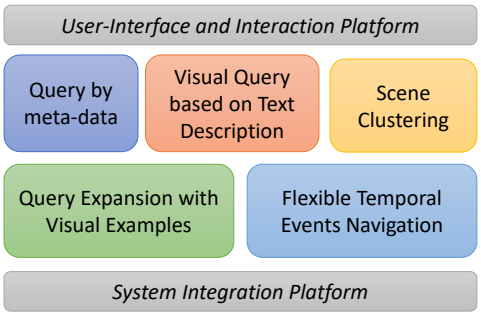

Figure 1: System overview of our query system.

We provide five main query processing components in our current system. Three components aim to retrieve photos related to certain criteria: Query by Meta-data, Visual Query based on Text Description, and Scene Clustering. We describe these modules in Section 3.2, 3.3, and 3.4, respectively. Our system also allows users to combine different functions to refine their search results.

We develop two additional functions to assist users and Query Expansion and Event Navigation and Exploration, which are presented in Section 3.6 and 3.5. These two features use an initial candidate photo as the seed point for expanding the query results, verifying the candidate with past or future events, etc.

We also briefly present the architecture for AI analysis of lifelog data and AI service pipeline definition in Sections 3.7 and 3.8.

3.2 Query by Meta-Data

Besides photos or video clips, lifelog data may contain additional data from different sources, e.g. GPS, location, time, personal health data, etc. Such data is valuable to narrow down the search space for the moments of interest. We support the hierarchical relationships between locations, such as *Helix Cafe* in *Dublin*, and *Dublin* is in *Ireland*, etc. We also support searching based on date and time range, or day in week. Figure 2 presents the form to query based on location and time.

In our system, we further analyze to extract enrich meta-data with semantic concepts from photos. We use scene text [22] to augment information for objects and places, such as the brand name of a product, a number appearing in the scene, or the name of a place.

We extract both general and personalized objects [16, 17]. General entities can be detected with existing pre-trained object detectors, such as Faster RCNN[26] or EfficientDet[28]. We also train our own object detectors to localize items that are unknown to generic object detectors but usually appear in personal activities, such as *coffee machine* or *medicine cupboard*.

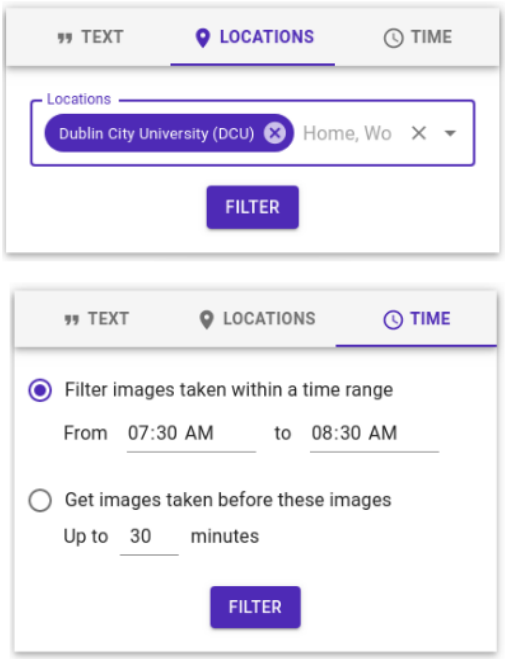

Figure 2: Query by location and time.

3.3 Text Query and Visual Information Matching based on Joint Embedding Model

We employ our Self-Attention based Joint Embedding Model (SAJEM [31]) in our retrieval system FIRST version 2. As illustrated in Figure 3, there are two branches to process visual and text data.

The image branch (in the left) detects regions in an input image and learns the interactions between those regions using our proposed Self-Attention Module [31] to create the representation of the image.

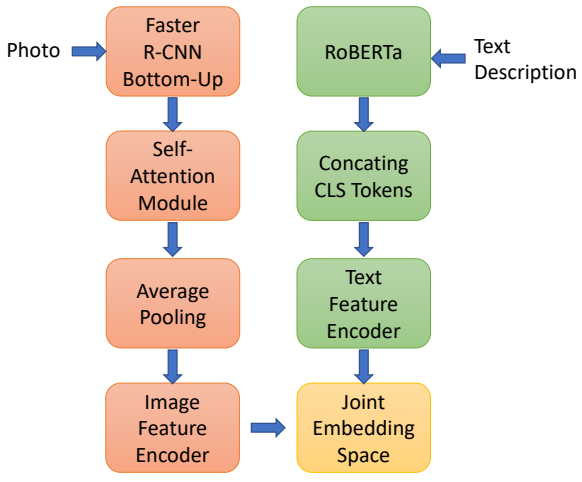

Figure 3: Main steps for Joint Embedding Model for visual and text feature comparison.

Based on the Multi-Head Self-Attention from Transformer mode [32], we replace Scaled Dot-Product Attention with Dot-Product Attention and use only one head to realize our image branch. For the text branch (in the right), we adopt RoBERTa [23] to encode a text description.

For joint embedding learning, we extract feature vectors of an image and its corresponding caption using the mentioned-above models, then feed these vectors into Image and Text Feature Encoders, respectively. In this way, we can project feature vectors of images and captions into a common space.

3.4 Scene Clustering based on Visual and Location Information

Clustering images help users to handle images in groups easily. As mentioned in [30], our system aims to support different criteria for image clustering. Currently, we implement two common strategies: grouping images based on their visual similarity and location.

Figure 4 demonstrates the visualization of image clusters based on their visual similarities (with visual features extracted by ResNet152). For photo clusters based on geolocation, we visualize them on map so that users can perform query based on locations.

Figure 4: Visualization of image clusters[30].

3.5 Flexible Temporal Events Navigation

To find or verify a certain event, we may need to look backward or forward a starting time instant. This idea was proposed and implemented in our previous systems, such as LifeSeeker 2.0[18] or FIRST 1.0 [30]. From a seed point, an initial photo corresponding to some criteria in the query, we can expand the sequence of photos to freely navigate in the timeline to refine the result, or to check for another event happening before or after the initial event.

As users may need to navigate slowly or quickly across time from a photo, our system supports users to adjust the time step interval, the granularity of the temporal data, to visualize the photos at specified intervals using the time slider. The step interval varies from seconds to hours, even across days, depending on the situation.

Figure 5 illustrates an example of sequence navigation for event verification. We can easily retrieve candidate photos corresponding to *in bus* environment, then expand the sequence of photos surrounding an initial photo to check the destination of that bus trip.

Figure 5: Example of sequence navigation for event verification.

3.6 Query Expansion with Visual Examples

To query similar images, we use ResNet152 features extracted from an initial photo and other photos in the dataset. Each photo is represented as a 2048-dimension feature. To speed up the process, we pre-calculate and store visual features of all photos in the collection. To evaluate the similarity between photos, we use the cosine distance between their features.

3.7 Service Integration for Lifelog Data Analysis

Because lifelog data has a large volume of data and needs many different processing tasks on it, we propose and build a solution using Google Colab and flexibly integrate Lifelog data analysis services. With built-in service integration, our system allows to change or add new processing features quickly by replacing and assembling processing modules through exporting results from API services. As needs change, systems and services can change according to needs.

Google Colab is a suitable choice for us to build data processing and analysis services. Because it provides servers with integrated GPUs for powerful modeling capabilities and rapid compatibility with cloud storage like Google Drive to leverage storage, security, and permission.

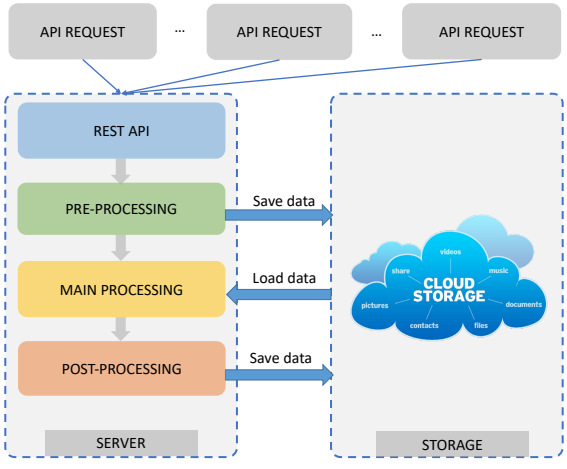

Figure 6: Flexible data analysis service for lifelog data.

Various lifelog data processing services such as object detection, place detection, image captioning, etc., can be implemented and deployed on different Google Colab instances as different servers. Each serving instance handles a particular analysis and exports the results to the outside via the API. Services all share the same process, including pre-processing, main processing, and post-processing in Figure 6. Pre-processing is responsible for receiving input from the API requests then normalizing the input for analysis. Main processing is the main processing function that uses algorithms and models depending on user needs. And post-processing is the part that takes care of saving the processing results to cloud storage. The analytics services are then treated as separate APIs on instances of Google Colab. When the system needs to process lifelog data according to one or more different types of analysis such as object detection, place detection, image captioning, etc., it only needs to call the right APIs.

3.8 Forming a Pipeline by Linking Components together Using NodeRed

Due to the nature of the solution where data is processed in a diverse, highly componentized environment, we propose a method to utilize existing technologies to accelerate the development process, while maintaining the final quality. Without having to re-implement existing functionalities, more time can be spent into research and development for critical components. Furthermore, existing tools are independently developed and thus, its stability is guaranteed. NodeRED is chosen for this purpose, because it provides a decent platform to design data flows, as well as orchestrating the execution.

On the front-end, NodeRED's user interface is intuitive for even users without a technical background. NodeRED uses the blackbox approach, where each component that represents a function are represented as nodes on the design surface. Each nodes have its own sets of input and output where data is received and sent, respectively. Nodes of functionality are laid out by dragging and dropping from the available nodes. The flow from the source to destination can be made by connecting inputs and outputs of nodes together. This therefore, establishes an execution graph that is easily understood by both humans and machines.

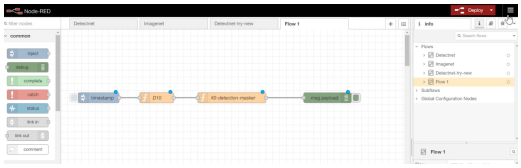

Figure 7: NodeRED design surface, with nodes connected to each other.

On the back-end, NodeRED uses NodeJS, a popular open-source, cross-platform, JavaScript runtime environment. Therefore, creating a new node for NodeRED is simple and standardized as creating a module for NodeJS, which opens the possibilities for automated updates and programmability.

4 SOME USAGE SCENARIOS

Description: Following someone's red backpack after having a coffee in Angelina's Cafe on a cold day...

We can first search for photos of *have coffee at Angelina's Cafe*, then look for red backpack in sequences following that moment. However, we think that looking for red backpack can help us narrow down the search space. Figure 8 shows the screenshots from our system to search for photos with *red backpack*. From this initial candidate photos, we use backward navigation to check the event of *having coffee at Angelina's Cafe*.

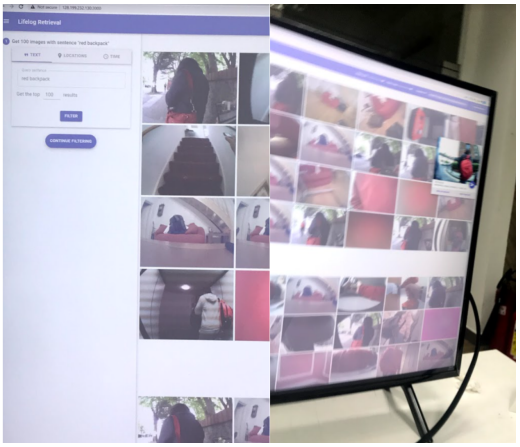

Figure 8: Search for *red backpack*.

Description: Buy a salt lamp in a shop.

First we may think to search for the concept *lamp*, but actually there is no such salt lamp among the candidate images. Then, we search for *shopping center*, and receive a long list of photos. By using scene text, recognized from photo, we can now find that moment, as represented in Figure 9.

Figure 9: Search for *salt lamp* with concepts extracted with scene text.

Description: Looking at ancient Chinese vases in a museum. There were two of them. They were blue and white in a wood and glass case.

For this query, we search for *blue vase* and further explore photos before and after that moment to confirm the location to be a museum. As shown in Figure 10, and can easily find the photo with two vases. From that photo, we can *View Timeline* to explore photos before and after it.

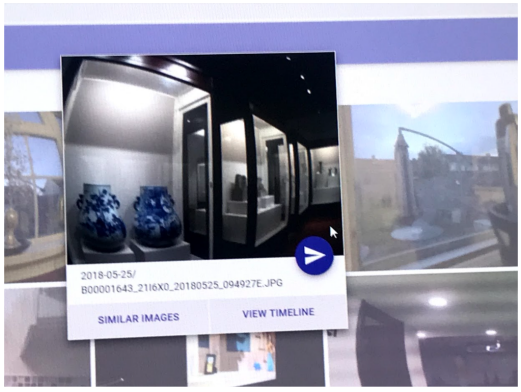

Figure 10: Search for *blue vase in museum.*

Description: Someone was looking inside the medicine cabinet in the bathroom at home.

We use the query (in text) "looking inside the medicine cabinet in the bathroom" to find some initial result, then we find similar images (query expansion) with ResNet152 feature to find remaining images. The query expansion function is useful to search for all events corresponding to a query (Figure 11).

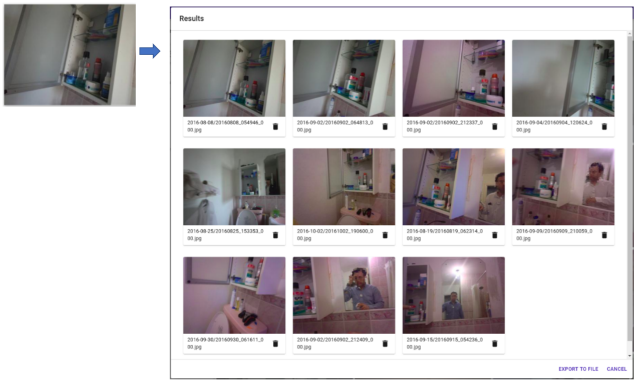

Figure 11: Use query expansion to find all events corresponding to a query.

Description: A woman in red top standing in front of a poster.

We find one image when searching with the query (in text) *a woman in red top standing in front of a poster*. To find another result, we perform the two-step search. First we retrieve top 1000 images with the query *a woman in red*, and refine the result with another query *poster on the wall*. Figure 12 shows the two photos corresponding to this query.

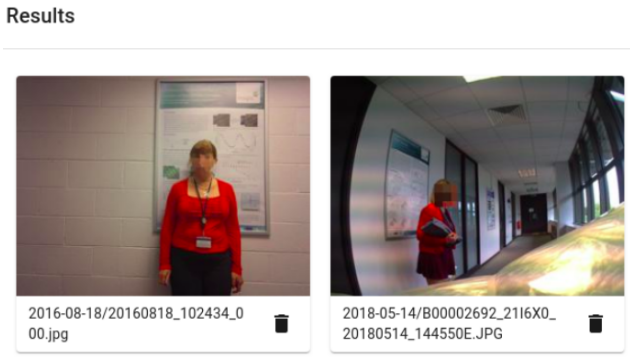

Figure 12: Iterative query with multiple steps.

5 CONCLUSION AND FUTURE WORK

In this paper, we present our interactive retrieval system for lifelog exploration. This is the second version of our flexible retrieval platform FIRST[30]. Based on the user-interface and interactive platform, we can easily build various UI forms for users to interact with. With our system integration platform, we can add more query processing modules to our solution. Currently, our system consists of five main modules to handle query with meta-data, query with text description, scene clustering, query expansion, and temporal navigation.

To handle more flexible ways for user to interact with, our system needs to be gradually upgraded new functions and modalities that we can learn from other research teams and solutions. Furthermore, we intend to evaluate the efficiency of different interaction and query approaches to enhance the ease-of-use for our system.

ACKNOWLEDGMENTS

This research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19.

Hoang-Phuc Trang-Trung, Thanh-Cong Le, and Mai-Khiem Tran were funded by Vingroup Joint Stock Company and supported by the Domestic Master/ PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Vingroup Big Data Institute (VINBIGDATA), code VINIF.2020.ThS.JVN.03, VINIF.2020.ThS.JVN.05, and VINIF.2020.ThS.JVN.06, respectively.

REFERENCES

[1] 2018. *LSC '18: Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge* (Yokohama, Japan). Association for Computing Machinery, New York, NY, USA.

[2] 2019. *LSC '19: Proceedings of the ACM Workshop on Lifelog Search Challenge* (Ottawa ON, Canada). Association for Computing Machinery, New York, NY, USA.

[3] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2020. Voxento: A Prototype Voice-Controlled Interactive Search Engine for Lifelogs. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge* (Dublin, Ireland) *(LSC '20)*. Association for Computing Machinery, New York, NY, USA, 77–81. https://doi.org/10.1145/3379172.3391728

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX (Lecture Notes in Computer Science)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.), Vol. 12375. Springer, 104–120. https://doi.org/10.1007/978-3-030-58577-8_7

[5] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8415–8424.

[6] 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.

[7] Aaron Duane, Björn Þór Jónsson, and Cathal Gurrin. 2020. VRLE: Lifelog Interaction Prototype in Virtual Reality: Lifelog Search Challenge at ACM ICMR 2020. Association for Computing Machinery, New York, NY, USA.

[8] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 12. http://bmvc2018.org/contents/papers/0344.pdf

[9] 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.

[10] 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.

[11] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Liting Zhou, Aaron Duane, Andreas Leibetseder, Michael Riegler, Luca Piras, Minh-Triet Tran, Jakub Lokoč, and Wolfgang Hürst. 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.

[12] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2021. Introduction to the Fourth Annual Lifelog Search Challenge, LSC'21. In *Proc. International Conference on Multimedia Retrieval (ICMR'21)*. ACM, Taipei, Taiwan.

[13] Silvan Heller, Mahnaz Amiri Parian, Ralph Gasser, Loris Sauter, and Heiko Schuldt. 2020. Interactive Lifelog Retrieval with Vitrivr. Association for Computing Machinery, New York, NY, USA.

[14] Omar Shahbaz Khan, Mathias Dybkjær Larsen, Liam Alex Sonto Poulsen, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, Dennis Koelma, and Marcel Worring. 2020. Exquisitor at the Lifelog Search Challenge 2020. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge* (Dublin, Ireland) *(LSC '20)*. Association for Computing Machinery, New York, NY, USA, 19–22. https://doi.org/10.1145/3379172.3391718

[15] Gregor Kovalcík, Vít Skrhak, Tomás Soucek, and Jakub Lokoc. 2020. VIRET Tool with Advanced Visual Browsing and Feedback. In *Proceedings of the Third ACM Workshop on Lifelog Search Challenge, LSC@ICMR 2020, Dublin, Ireland, June 8-11, 2020*, Cathal Gurrin, Klaus Schöffmann, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoc, Minh-Triet Tran, and Wolfgang Hürst (Eds.). ACM, 63–66. https://doi.org/10.1145/3379172.3391725

[16] Nguyen-Khang Le, Dieu-Hien Nguyen, Trung-Hieu Hoang, Thanh-An Nguyen, Thanh-Dat Truong, Tung Dinh Duy, 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, LSC@ICMR 2019, Ottawa, ON, Canada, 10 June 2019*, Cathal Gurrin, Klaus Schöffmann, Hideo Joho, Duc-Tien Dang-Nguyen, Michael Riegler, and Luca Piras (Eds.). ACM, 1–6.

[17] Nguyen-Khang Le, Dieu-Hien Nguyen, Vinh-Tiep Nguyen, and Minh-Triet Tran. 2019. Lifelog Moment Retrieval with Advanced Semantic Extraction and Flexible Moment Visualization for Exploration. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019 (CEUR Workshop Proceedings)*, Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller (Eds.), Vol. 2380. CEUR-WS.org. http://ceur-ws.org/Vol-2380/paper_139.pdf

[18] 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. Association for Computing Machinery, New York, NY, USA.

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV (Lecture Notes in Computer Science)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11208. Springer, 212–228. https://doi.org/10.1007/978-3-030-01225-0_13

[20] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual Semantic Reasoning for Image-Text Matching. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 4653–4661. https://doi.org/10.1109/ICCV.2019.00475

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX (Lecture Notes in Computer Science)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.), Vol. 12375. Springer, 121–137. https://doi.org/10.1007/978-3-030-58577-8_8

[22] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. 2020. ABCNet: Real-Time Scene Text Spotting With Adaptive Bezier-Curve Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[24] František Mejzlík, Patrik Veselý, Miroslav Kratochvíl, Tomáš Souček, and Jakub Lokoč. 2020. SOMHunter for Lifelog Search. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge* (Dublin, Ireland) *(LSC '20)*. Association for Computing Machinery, New York, NY, USA, 73–75. https://doi.org/10.1145/3379172.3391727

[25] Van-Tu Ninh, Tu-Khiem Le, Liting Zhou, Luca Piras, Michael Riegler, Pål Halvorsen, Mathias Lux, Minh-Triet Tran, Cathal Gurrin, and Duc-Tien Dang-Nguyen. 2020. Overview of ImageCLEF Lifelog 2020: Lifelog Moment Retrieval and Sport Performance Lifelog. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020 (CEUR Workshop Proceedings)*, Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol (Eds.), Vol. 2696. CEUR-WS.org. http://ceur-ws.org/Vol-2696/paper_65.pdf

[26] 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

[27] Luca Rossetto, Matthias Baumgartner, Narges Ashena, Florian Ruosch, Romana Pernischová, and Abraham Bernstein. 2020. LifeGraph: A Knowledge Graph for Lifelogs. In *Proceedings of the Third ACM Workshop on Lifelog Search Challenge, LSC@ICMR 2020, Dublin, Ireland, June 8-11, 2020*, Cathal Gurrin, Klaus Schöffmann, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoc, Minh-Triet Tran, and Wolfgang Hürst (Eds.). ACM, 13–17. https://doi.org/10.1145/3379172.3391717

[28] Mingxing Tan, Ruoming Pang, and Quoc V. Le. 2020. EfficientDet: Scalable and Efficient Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10778–10787.

[29] 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* (Dublin, Ireland) *(LSC '20)*. Association for Computing Machinery, New York, NY, USA, 23–28. https://doi.org/10.1145/3379172.3391719

[30] Minh-Triet Tran, Thanh-An Nguyen, Quoc-Cuong Tran, Mai-Khiem Tran, Khanh Nguyen, Van-Tu Ninh, Tu-Khiem Le, Hoang-Phuc Trang-Trung, Hoang-Anh Le, Hai-Dang Nguyen, Trong-Le Do, Viet-Khoa Vo-Ho, and Cathal Gurrin. 2020. FIRST - Flexible Interactive Retrieval SysTem for Visual Lifelog Exploration at LSC 2020. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge* (Dublin, Ireland) *(LSC '20)*. Association for Computing Machinery, New York, NY, USA, 67–72. https://doi.org/10.1145/3379172.3391726

[31] Hoang-Phuc Trang-Trung, Hoang-Anh Le, and Minh-Triet Tran. 2020. Lifelog Moment Retrieval with Self-Attention based Joint Embedding Model. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020 (CEUR Workshop Proceedings)*, Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol (Eds.), Vol. 2696. CEUR-WS.org. http://ceur-ws.org/Vol-2696/paper_60.pdf

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.