An Interactive Multimodal Retrieval System for Memory Assistant and Life Organized Support

Van-Luon Tran
University of Science, VNU-HCMC
Ho Chi Minh, Vietnam
1612362@student.hcmus.edu.vn

Anh-Vu Mai-Nguyen
University of Science, VNU-HCMC
Ho Chi Minh, Vietnam
1612904@student.hcmus.edu.vn

Trong-Dat Phan
University of Science, VNU-HCMC
Ho Chi Minh, Vietnam
1512102@student.hcmus.edu.vn

Anh-Khoa Vo
University of Science, VNU-HCMC
Ho Chi Minh, Vietnam
1512262@student.hcmus.edu.vn

Minh-Son Dao
NICT
Tokyo, Japan
dao@nict.go.jp

Koji Zettsu
NICT
Tokyo, Japan
zettsu@nict.go.jp

ABSTRACT

Lifelogging is known as the new trend of writing diary digitally where both the surrounding environment and personal physiological data and cognition are collected at the same time under the first perspective. Exploring and exploiting these lifelog (i.e., data created by lifelogging) can provide useful insights for human beings, including healthcare, work, entertainment, and family, to name a few. Unfortunately, having a valuable tool working on lifelog to discover these insights is still a tough challenge. To meet this requirement, we introduce an interactive multimodal retrieval system that aims to provide people with two functions, memory assistant and life organized support, with a friendly and easy-to-use web UI. The output of the former function is a video with footages expressing all instances of events people want to recall. The latter function generates a statistical report of each event so that people can have more information to balance their lifestyle. The system relies on two major algorithms that try to match keywords/phrases to images and to run a cluster-based query using a watershed-based approach.

CCS CONCEPTS

• **Human-centered computing** → **User centered design**; User centered design; • **Information systems** → **Top-k retrieval in databases**; **Image search**; *Top-k retrieval in databases*; • **Computer systems organization** → Client-server architectures; • **Theory of computation** → *Unsupervised learning and clustering*; *Semi-supervised learning*; • **Computing methodologies** → **Object detection**; **Information extraction**; **Neural networks**.

KEYWORDS

Lifelogs, clustering, image retrieval, feature extracting, content and context, semantic

ICMR '20, June 8–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06…$15.00
https://doi.org/10.1145/3372278.3391934

ACM Reference Format:
Van-Luon Tran, Anh-Vu Mai-Nguyen, Trong-Dat Phan, Anh-Khoa Vo, Minh-Son Dao, and Koji Zettsu. 2020. An Interactive Multimodal Retrieval System for Memory Assistant and Life Organized Support. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3372278.3391934

1 INTRODUCTION

Recently, the terminologies of lifelogging and lifelog became familiar to researchers from various domains [3]. The former expresses the action of continuously recording data using wearable sensors to reflect the world under the first viewpoint. The latter denotes the data collected by lifelogging. In other words, people treat lifelog as their diary where, in the future, they can look back and understand more about how their lives are going. Various exciting information extracted from lifelogging can become a reliable and objective source for understanding different aspects of people's lives, such as healthcare, daily activities, consuetudines, entertainment, and education.

Nevertheless, the variety (e.g., images, physiological data, tags), volume (e.g., daily-captured data, high-resolution photos), and veracity (e.g., subjective emotion, noise) of lifelog make knowledge discovery complicated. These troublesome issues lead to the emerging requirements of having a solution to structure, process, understand, and query lifelog. Several exciting challenges have been introduced to communities to solve different problems related to lifelog towards contributing and promoting the research of lifelog, such as NCTIR [1], ImageCLEFlifelog [2], and Lifelog Search Challenge (LSC) [3].

Along with this trend, we build an insights-from-lifelog system that can make life easier by offering the memory assistant and life organized support functions. The former aims to list all instances of the event whose content is simply sketched by users (e.g., a few keywords that are pop-up from users' minds related to the moments they want to recall), under a video with footage format. The latter targets to create a statistical report of a set of specific activities so that they can readjust their life upon that (e.g., users can have a summary of the portion between their business trip and family

[1] http://research.nii.ac.jp/ntcir/index-en.html
[2] https://www.imageclef.org/2020/lifelog
[3] http://lsc.dcu.ie/

time so that they can rearrange their lifestyle better). The core of this system is the interactive multimodal retrieval framework that utilizes several deep learning models running on textual, visual, and signal modalities to query events. Besides, the friendly Web-UI and enrich query content modules are other advantages of this system when offering users a powerful tool to polish their queries and results.

This paper is organized as follows: Section 2 reviews related works for a similar problem, Section 3 explains the system, Section 4 introduces two examples of how to use our system, and Section 5 gives conclusions and future works. Our contribution can be focused on two points:

- The attention mechanism processes the text-based query. Instead of hard matching between individual text query and image (i.e., comparing classes of objects which appear in image with keywords in the input text), we try to learn the relationship between two spaces (text space and visual space) and match each item from one space to suitable item in another space by using the probability that indicates how relevant they are.
- A friendly UI web application allows users to interact with our system. Our motivation is not only to let users manipulate with fewer steps as possible but also to ensure the best result they can have.

2 RELATED WORK

Interactive systems for lifelog exploring and exploiting have attracted many researchers. In [2], the authors develop a Virtual Reality Lifelog Explorer system focusing on utilizing a gesture-based interactive manner to make people comfortable. First, a user describes his/her queries by selecting suitable tags by virtually touching two submenus time and concepts. After the user submits his/her query, the system looks for a result and displays that result in decreasing tag-rank order (i.e., the more number of tags, the higher ranking). Another interactive system, called LIFER, is introduced in [8]. Similar to the VRLE system, LIFER is a facet search engine that allows users to choose the keywords corresponding to their queries by using facet filters. An image containing more keywords will get a higher score. In [5], the lifelog retrieval system, namely LifeSeeker, is introduced. This system provides users with two-ways for querying: free-text and image similarity. For the former, they apply the term-weighting method [7] to rank images. For the latter, they compare the histogram between the query image and dataset. Finally, the top 100 images (configurable) are displayed.

3 INTERACTIVE MULTIMODAL RETRIEVAL SYSTEM

In this section, we introduce the architecture, data, core, and user interface of the system.

3.1 System Architecture

In this subsection, we introduce the client-server architecture that is depicted in Figure 1, applying it in our retrieval system.

- **Client**: We design a web UI application named Retrieval Application to provide a friendly user interface. The Retrieval Application allows users to submit their query (text or image) and other particular conditions (via filters), to modify returned output and to export a final record. In this way, users can probably get the most satisfying result for their query.
- **Server**: We introduce two types of servers; each of them plays a different role in our system.
  - *Application server*: Firstly, the Application server receives a query from users and converts it into a suitable form, which is later used as an input for Core Server. Then, the Application server waits for Core server responses, handles indexes of images returned by the Core server. Finally, it stores these indexes in the Cache, applies filters, and gets corresponding images (pointed by these indexes mentioned above) from the Storage.
  - *Core server*: We implement a deep learning algorithm that includes indexing, searching, and querying functions, and returns raw results, which is modifiable by users at Retrieval Application.

  Alongside main components (application and server), we also design auxiliary parts for making our system work efficiently.
  - *Cache*: Temporarily saves indexes of images for later uses
  - *Storage*: Contains images and their metadata from provided dataset.

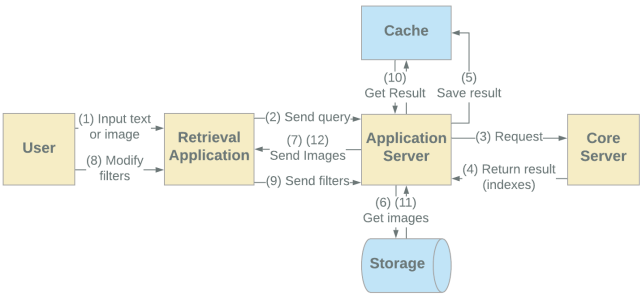

Figure 1: System overview

3.2 Data

We use the data provided by Lifelog Search Challenge 2020 [4]. This data is the combination of NTCIR Lifelog datasets 2015, 2016, and 2018. Over 191.000 images have been recorded for 114 days. Moreover, the dataset contains metadata including time, hearts, steps, and calorie data collected from other sensors. It should be noted that metadata and images are synchronized by the time, and not all images have their associated metadata.

3.3 Modeling

Our system relies upon two core algorithms as described below:

- *Text-to-sample module*: From the input query, we applied the attention mechanism described in [1] to first recognize keywords in the sentence, then to match these keywords with relevant samples. We define a sample as an image that is the

[4] http://lsc.dcu.ie/

best visual representation of the semantic meaning implied from these keywords (i.e., word-2-image). The purpose of the attention mechanism here is to map words or phrases with their most similar images (usually one-by-one) and vice versa.

- *Query module*: The Faster-RCNN (backbone ResNet-101) is utilized to vectorize both samples and a whole dataset. In our experiment, we firstly extract 2048-D vector for each instance (i.e., image or objects inside an image). Then, we find all vectors in the dataset which are the same as our samples' vectors. We use cosine similarity as our criterion and gather vectors into a database after embedding them on FAISS [4]. Moreover, clustering algorithms proposed in [6] are applied to support the searching process by grouping similar images together.

3.4 Web User Interfaces (Web UI)

Aiming to provide a friendly and convenient way to query data, we design our Web UI with the following functions.

3.4.1 Search bar. We allow users to create their queries with both text and image.

3.4.2 Filters. To enable the interactive retrieval schema, we provide users with a set of filters to let them polish their results as well as enrich their queries until they satisfy with the results. The purpose of these filters is to provide users a set of tags (i.e., keywords, phrases) that are generated from the database. Each tag can be seen as the weight to polish the results or narrow/broaden the scope of searching. When users click on any tag from a filter, the clicked tag is added to the query. Users can also remove any tag from the query by clicking on it. Figures 2 shows examples of using these filters. We create two filters that provide a name list of locations and objects associated with images stored in the database. The former is generated by utilizing Places365 model, and the latter is built with the deep-learning object detection model with 1500 classes. We also create other two filters for time and metadata (e.g., foot-steps, heart-beats, speed). The former lets users specialize in dates and times so that users can query by any period. The latter allows users to polish their results by additional information related to activities and physiological data.

3.4.3 Results Visualization. The retrieved result is displayed in this area. We visualize all found images in multiple-line where each line contains five images by default. It should be noted that we display these images by time direction. When users move a mouse over an image, a delete button, and information of this image (including time, location, and semantic information) are pop-up. Users can easily remove inaccurate images by clicking on the delete button. In that case, inaccurate images are still displayed on the interface and masked as deleted with a red overlay. Figures 3 illustrates the differences between original and modified results.

3.4.4 Data Exportation. Exporting feature is developed to serve a variety of purposes. After filtering all not relevant items, users can use the exporting feature to get the result as they want. The system provides an ability to export the results by different formats such as indexes, images, videos, or charts accordingly to users' desires.

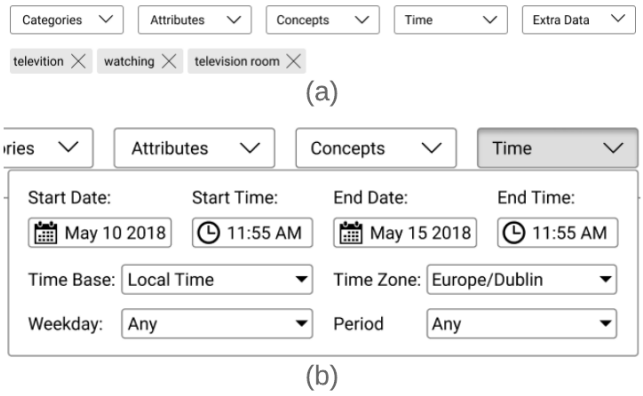

Figure 2: Filters

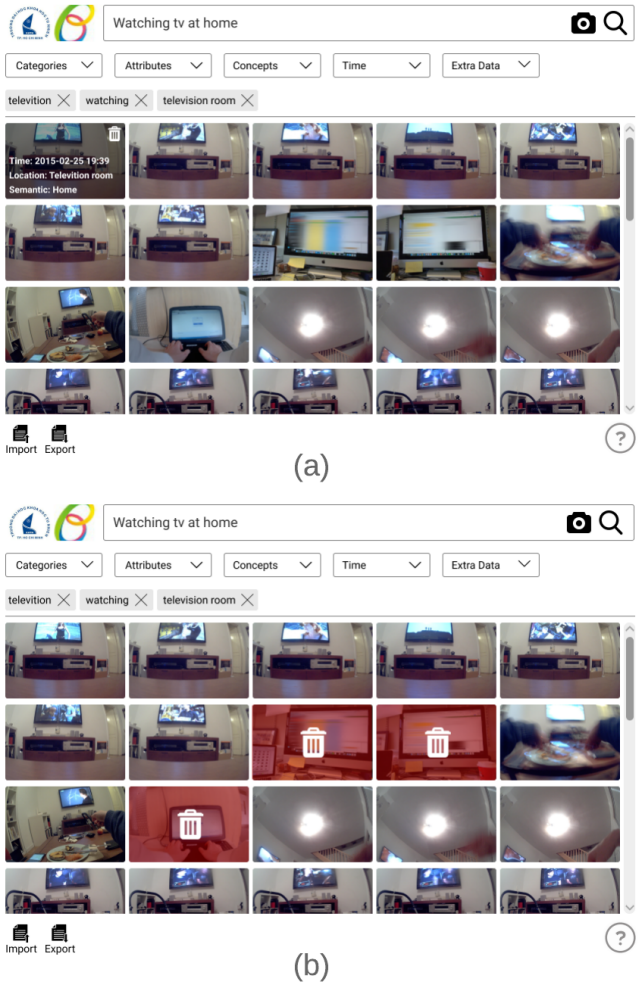

Figure 3: Results Polishing: (a) first result (b) polished result

Users can also import the indexes into our system to visualize associated images.

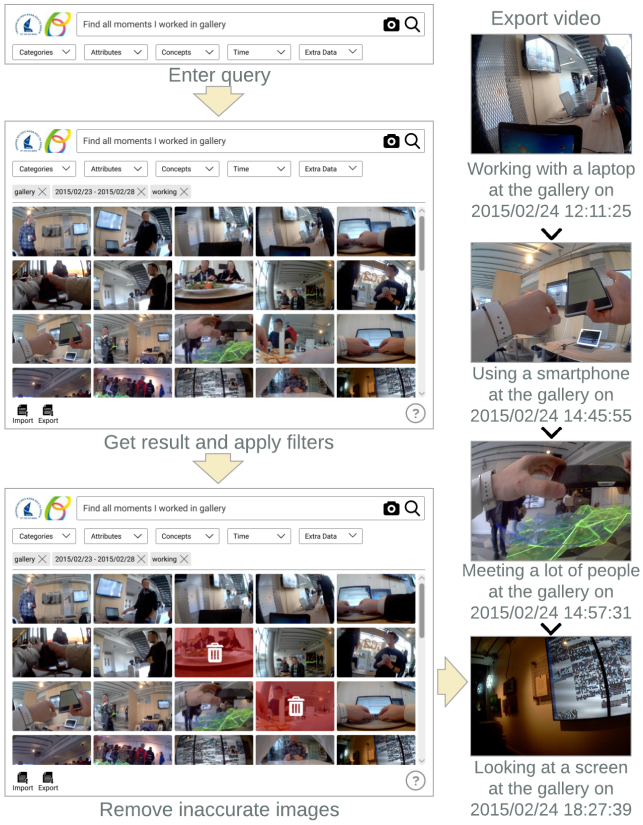

Figure 4: Memory Assistant Explication

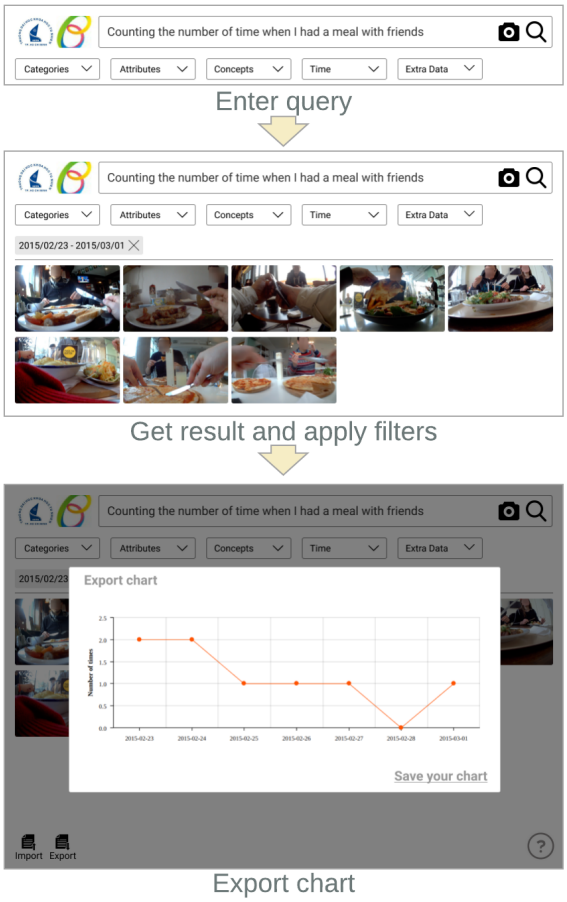

Figure 5: Life Organized Support Explication

3.4.5 Help. To help users quickly learn how to use the system, we provide the help function. We leverage the mouse hover manner to design this function so that whenever users move a mouse over components, related information is pop-up.

4 MEMORY ASSISTANT AND LIFE ORGANIZED SUPPORT

In this section, two case studies are introduced to explicate how our system can assist people in recalling a specific historical moment and provide them the necessary information to re-organize their lifestyle.

- *Memory Assistant*: One lifelogger has spent several times at one gallery. One day, the lifelogger wants to recall all his/her moments at that gallery exported as a video with footages (i.e., the annotation of each instance of the event generated automatically by our system). Our system can support this requirement by the query that looks like "find all the moments that I was at a gallery before March 2015." Figure 4 illustrates the interactive query progress and the exported data.
- *Life Organized Support*: Another lifelogger wants to know the balance between his/her working and free time in one month towards having a better lifestyle. Our system can help the lifelogger find all working and family/friend moments within one month, and export the statistical report to visualize the portion of these events. Figure 5 depicts the interactive query progress and the statistical chart. In this example, the user wants to know the report of "how many times I have had a meal with my friends during 2015/02/23-2015/03/01?"

5 CONCLUSION

We have introduced a new interactive multimodal retrieval system for discovering insights from lifelog. We briefly explain the algorithms that play as the search engine of the system. We also illustrate how to use the system to have a memory assistant and life organized support in a friendly and convenient manner with Web UI. In the future, more necessary functions, as well as more accuracy query algorithms, will be developed to bring more benefits to users who own a vast lifelog but just see the top of the ice-berg.

ACKNOWLEDGEMENT

This research is conducted under the Collaborative Research Agreement between National Institute of Information and Communications Technology and University of Science, Vietnam National University at Ho-Chi-Minh City.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[2] Aaron Duane, Cathal Gurrin, and Wolfgang Huerst. 2018. Virtual Reality Lifelog Explorer. In *Proceedings of the ACM Workshop on Lifelog Search Challenge*. 1–3.

[3] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. Lifelogging: Personal big data. In *Foundations and Trends in Information Retrieval*. 3–7.

[4] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734* (2017).

[5] Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Minh-Triet Tran, Liting Zhou, Pablo Redondo, Sinead Smyth, and Cathal Gurrin. 2019. LifeSeeker: Interactive Lifelog Search Engine at LSC 2019. In *Proceedings of the ACM Workshop on Lifelog Search Challenge*. 37–40.

[6] T. Phan, M. Dao, and K. Zettsu. 2019. An Interactive Watershed-Based Approach for Lifelog Moment Retrieval. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. 282–286. https://doi.org/10.1109/BigMM.2019.00-10

[7] S.E. Robertson and K. Sparck Jones. 1997. Simple, Proven Approaches to Text Retrieval. *Technical report* (1997).

[8] Liting Zhou, Zaher Hinbarji, Duc-Tien Dang-Nguyen, and Cathal Gurrin. 2018. LIFER: An Interactive Lifelog Retrieval System. In *Proceedings of the ACM Workshop on Lifelog Search Challenge*. 1–4.