HCMUS at the NTCIR-14 Lifelog-3 Task

Nguyen-Khang Le[1], Dieu-Hien Nguyen[1],
Trung-Hieu Hoang[1], Thanh-An Nguyen[1], Thanh-Dat Truong[1],
Duy-Tung Dinh[1], Quoc-An Luong[1], Viet-Khoa Vo-Ho[1],
Vinh-Tiep Nguyen[2], and Minh-Triet Tran[1]

[1] University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[2] University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam

Abstract. Lifelogging has been gaining more and more attention in the research community in recent years. Not only can it provide valuable insight and a deeper understanding of human daily activities, but it can also be used to improve personal health and wellness. However, there are many challenging problems in this field. One of the most important tasks of processing lifelog data is to access its semantic, which aims to retrieve the moments of interest from the lifelog. There are many approaches to this problem, two of which are using data processing and providing friendly user interaction. Our proposed system takes both of these approaches. We first extract concepts from the images, build a structure to quickly query images based on these concepts. We then provide users with a friendly user interface to perform the task.

Team Name. HCMUS

Subtasks. Lifelog Semantic Access subTask (LSAT) (English)

Keywords: Lifelog Retrieval · Habit-based concept detector · Moment sequence visualization.

1 Introduction

With the rise of technology over the last decade, the number of smart wearable devices, as well as low-cost sensors, have been increasing rapidly. These devices are more available than ever. Therefore, it allows anyone to use these devices to capture the details of their everyday's life and create an enormous dataset which can include photos, videos, biometric and GPS information. This type of dataset is commonly referred to as lifelog.

Dodge and Kitchin [4] refer to lifelogging as 'a form of pervasive computing, consisting of a unified digital record of the totality of an individual's experiences, captured multimodally through digital sensors and stored permanently as a personal multimedia archive'. This task was initially proposed because the organizers identified that technological progress had resulted in lifelogging becoming a normative activity, thereby necessitating the development of new forms

of personal data analytics and retrieval that are designed to operate on multimodal lifelog data [5]. Recent efforts to employ lifelogging has been noted, some as a mean of supporting human memory or facilitating large-scale epidemiological studies in healthcare, lifestyle monitoring diet/obesity or for exploring social issues such as privacy-related concerns or behavior analysis.

Lifelog analysis has a lot of benefits in research and applications, it can help to give a better intuition about human activities on a regular basis as well as improving their own wellness. Particularly, lifelogging analysis that aims [7] to retrieve the moments of interest from lifelog data can help people to revive memories [11], verify events, find entities, or analyze people's social traits [3]. There are many other challenging tasks in lifelog analysis that also have great potential in research and application. In this paper, we will focus on solving the Lifelog Semantic Access task (LSAT), whose mission is to retrieve moments of interest from the lifelog data.

The LSAT subtask was a known-item search task applied over lifelog data. In this subtask, the participants had to retrieve a number of specific moments in a lifelogger's life in response to a query topic. Moments are considered to be semantic events or activities that happened at least once in the dataset. The task can best be compared to a known-item search task with one (or more) relevant items per topic [5].

We view this problem as two separate subproblems. These two subproblems can be referred to as offline data processing and user interaction respectively.

- The first subproblem aims to preprocess the data and annotate each data with appropriate metadata. In this problem, we propose to optimize the metadata of the lifelog dataset, which includes places and concepts detected by a pre-trained detector, by adding a custom detector based on the user's habits. We also find an efficient way to index the lifelog data based on these metadata so that we can later quickly retrieve the data.
- The second subproblem aims to design and provide a friendly user interface that enables novice users to interact with the queries and select the result data. To solve this problem, we come up with a user interface design that makes it easy for users to input query information and maximize their ability to retrieve the correct moments from the query result.

From our first generation lifelog retrieval systems[13, 12], we create a new generation of retrieval systems focusing on user experience enhancement and video sequence exploration. With our system, we efficiently solved the 24 queries of Lifelog Semantic Access task (LSAT) of NTCIR14 and achieve the best result comparing to other runs in all metrics, including MAP, P@10, and RelRet.

In Section 2, we discuss some recent challenges and achievements in Lifelog research. We propose our methods in Section 3 where we focus on the offline data processing and user interaction. In Section 4, we give an example of how our system assists a novice user to retrieve the moments of interest from the lifelog. The conclusion and a discussion of what can be done in the future work are presented in Section 5.

2 Related Work

Comparing the performance of information access and retrieval systems that operate on lifelog data is among the interesting topics for researchers worldwide recently. One of the first significant conferences that focus on known-item search and activity understanding applied over lifelog data was NTCIR-12 which happened in 2016 [6]. The lifelog data used in this conference is collected from 3 different volunteers wearing cameras to record visual daily life data for a month. Furthermore, the conference also provides a concept detector to support the participating teams. Many different analytic approaches and applications are discussed in the conference due to the enormous amount of data in the lifelog.

The area of interest is widened to other aspects other than the origin information retrieval purpose [1][2]. In Lifelog Semantic Access Task challenge, some teams focus on offline data processing steps and created a retrieval system that runs in an automatic manner which does not require any user involved, while other teams try to optimize user interface design to better support novice users.

In ImageCLEFlifelog 2017, more information is added to the lifelog dataset, some are semantic locations such as coffee shops and restaurants, others are physical activities such as walking, cycling and running. The tasks on this dataset include a retrieval task which includes the evaluation of result image correctness, and a task in which the dataset is summarized by a specific requirement.

In Lifelog Search Challenge (LSC 2018), we proposed our very first system for interactive lifelog retrieval[13]. This system allows users to search for certain moments based on the concepts extracted from each image. In this system, there are 3 main components[13]: visual clustering, augmented data processing, concept extraction, and augmented data processing. Then we added the image captioning module into our system to further exploit the text representation of lifelog images [12] in CLEF 2018. Our first generation of lifelog retrieval systems provide users with four different types of queries on place, time, entity, and extra biometric data [13][7]. We also proposed an enhancement for image captioning [14] to improve the quality of generated captions.

Taking advantage of our previous works [13][12][14], we build a more efficient system that provides two main values: (i) a set of object detectors for detecting habit-based concepts and (ii) a friendly user interface supporting sequence view of images.

3 Proposed Retrieval System

3.1 Retrieval System Overview

To solve the Lifelog Semantic Access Task, we first evaluate the dataset to figure what kind of places and concepts we should focus on. After the evaluation, we find that the accurate and relevant information about the place and objects appearing on the lifelog data combined with a user interface that allows users to traverse back and forth from a specific moment will be sufficient to retrieve the moments of interest. Therefore, we break down the problem into the offline

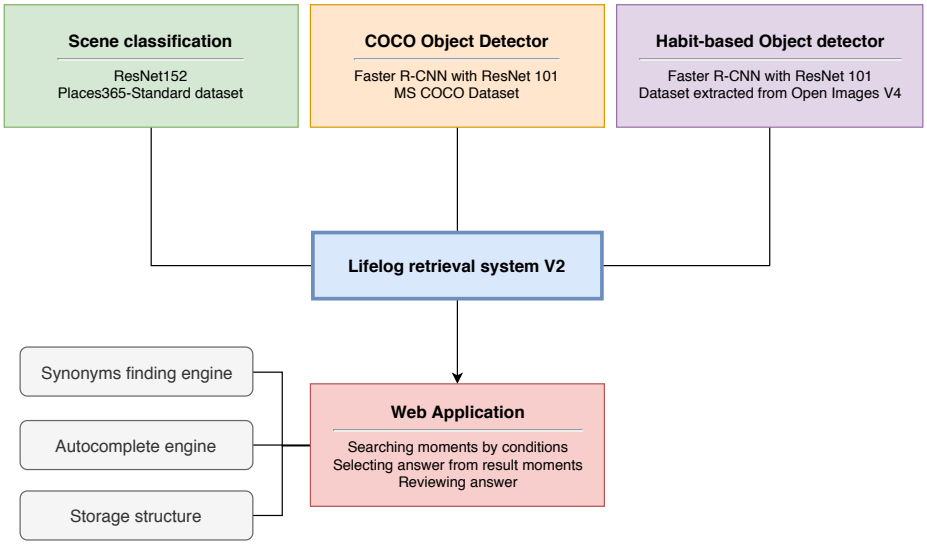

Fig. 1: Main components of the proposed system (Lifelog retrieval system V2)

data processing problem and user interaction problem. We then process to solve these problems separately.

Our system context diagram is demonstrate in Figure 1. The two main groups of components in our system are Offline Data Processing and User Interaction. Theses two main component groups are introduced in Sections 3.2 and 3.3 respectively.

In this paper, we focus on illustrating the usage of our retrieval system on solving a wide variation of queries in LSAT in NTCIR14. For the design and a detail description of different components in our system, please refer to [9].

3.2 Offline data processing

In the offline data processing step, we have two main goals. First, we aim to annotate each image in the lifelog with the metadata that consists of the information about scene's category, scene's attributes, and appearing concepts. Second, we provide a method to index the dataset based on these metadata for fast retrieval.

For object detection, we employ a basic object detector trained on MS COCO 2014 dataset [10] and our habit-based detectors that take advantage of the Open Image V4 dataset [8]. All of the above detectors use Faster R-CNN with ResNet 101 backbone. Furthermore, a classifying model is trained to predict the scene's category and scene's attributes. For this classifying model, ResNet 152 is employed and trained on Places365-Standard dataset [15]. Finally, we propose a storing structure which utilizes hash table and tree to make the retrieving process more efficient.

3.3 User interaction

A friendly user interface is one of the most important aspects of our system. This user interface design must meet these two goals:

1. The novice user can easily query the images from the dataset with the desired attributes.
2. The novice user can traverse back and forth from a specific result moment, and choose what images are the correct ones.

We design a friendly user interface that meets these goals and develop a web application applying this design. In our web application, we provide the user with 3 main views, the user can easily switch between these views:Search mode, Result mode, and Semi mode.

Moreover, a view for reviewing the answers is also provided. In this view, the user can view all of his/her chosen images for a specific topic, remove incorrect ones and change the images' order.

4 Experiment with Queries in LSAT of NTCIR14

4.1 Overall

In this section, we present how our system performs in practice where it is used to retrieve the moments in the lifelog which corresponds to a given search topic. The system automatically generates some of the input fields for the user and allows the user to modify them to get the correct result. The system provides a flexible way and multiple tools for the user to do the task, but the user also needs to picture the moments and decide what needs to be in the inputs in order to get a more precise result. Although our system's user interface is user-friendly and self-explained itself, many tooltips and pop up instructions are supported to guide the user.

4.2 Dataset

The detail of the dataset gathering process is described in [5]. The lifeloggers wore an OMG Autographer passive-capture wearable camera clipped to clothing or worn on a lanyard around the neck which captured images (from the wearer's viewpoint) and operated for 12-14 hours per day (1,250 - 4,500 images per day - depending on capture frequency or length of waking day). Additionally, the lifeloggers included locations and physical movements and a record of music listening. Finally, the dataset included health and wellness data from continual heart-rate monitors, continuous (15-minute interval) blood glucose monitors, along with manual annotations of food and drink consumption.

4.3 Search topic

Find the moment when u1 was eating ice-cream beside the sea.

NOTE: *To be relevant, the moment must show both the ice-cream with a cone in the hand of u1 as well as the sea clearly visible. Any moments by the sea, or eating ice-cream which do not occur together are not considered to be relevant.*

At first glance, it was clear that the scene's category of this topic was the beach and the main concept was an ice-cream. There are multiple approaches to this topic.

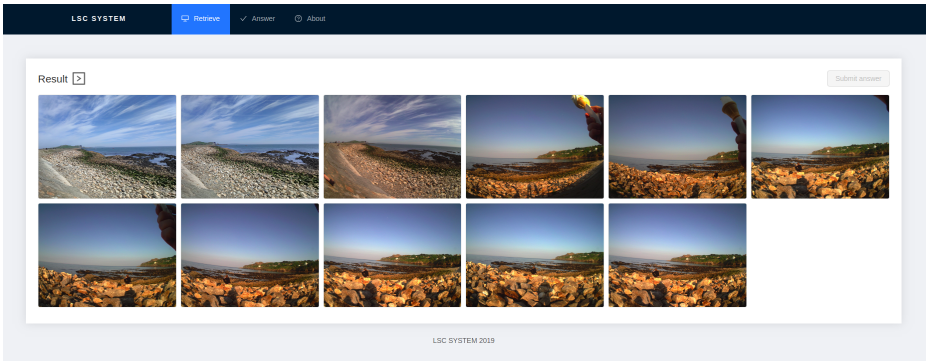

Fig. 2: Result images when querying for scene's category "coast"

In the first approach, we did not use the information of the ice-cream but just the scene information which was "beside the sea". Our first guess was that this scene's category should be "beach". When we input the word "beach" in the scene's category field, our system automatically found all synonyms and related words of this word to see if any of these words match with our scene's categories or concepts. As a result, a synonym of "beach" which is "coast" was found and automatically fill into the input field. The system then retrieved all the images from the lifelog dataset that match the query (Figure 2). From here, we could easily pick the correct moment when the lifelogger was eating ice-cream beside the sea. With the help of image sequence view, we were able to retrieve all the correct images taken around that time whether they contained the beach or not.

In the second approach, we focused on the concept of ice-cream and did not take advantage of the scene information. With the habit-based detector, our system was able to detect the concept ice-cream which is one of the concepts in our Dessert detector (Figure 3). As a result, our system correctly retrieved all the images with ice-cream appearing. Similar to the previous approach, with the help of image sequence view, all correct images were chosen.

Finally, we used the review answer page to finalize our result, including removing unwanted images and changing the order of the images to get the best mean average precision result (Figure 4).

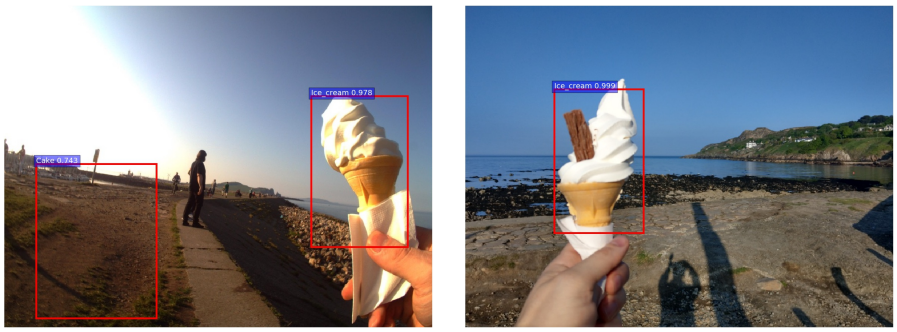

Fig. 3: Detecting results of concept "Ice cream" using habit-based detector

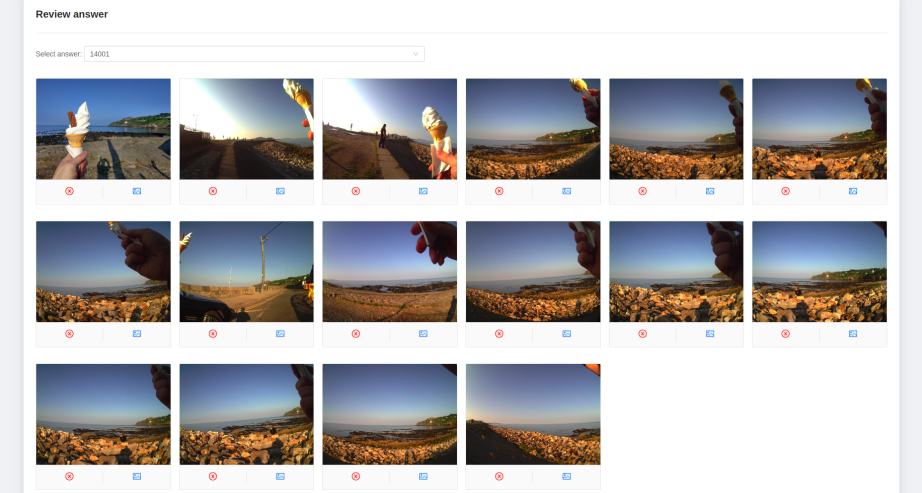

Fig. 4: Reviewing answer for topic: "Find the moment when u1 was eating ice-cream beside the sea."

Find the moment when u1 was eating fast food alone in a restaurant.

NOTE: *U1 was eating fast food in a restaurant while away from home. He ate chips and a hamburger with a drink. Moments when eating hamburgers cooked at home on a BBQ are not relevant.*

The scene's category was a restaurant and the main concept was fast food. First, we tried to query for the scene's category "fast food restaurant" as this is one of the categories in the Places365 standard dataset, the system gave us

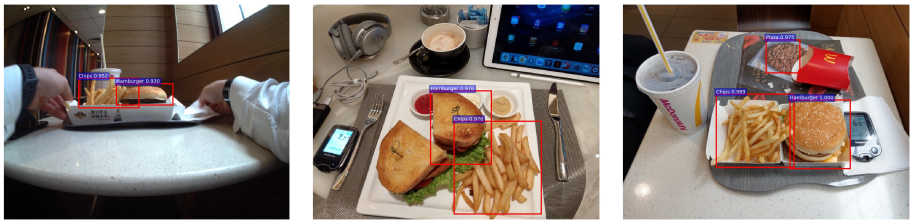

Fig. 5: Detecting results of one of our habit-based detectors relating to food

a lot of result images, some of which were not fast food restaurants but coffee shops. The system did not do a good job as classifying fast food restaurant. Thus, we searched the concept of fast food. The system found some synonyms of fast food such as "convenience food" and "junk food", none of which belongs to our concept's categories so we have to guess the fast food ourselves.

We first thought about hamburger as common fast food and made the query for this concept. The MS COCO 2014 dataset does not have the hamburger object. However, the detector for "Main food", one of our habit-based detectors does have this concept so our system was able to retrieve images with hamburgers. After the query, plenty of the result images did have hamburgers but were not exactly in a fast food restaurant (Figure 5). We then used the image sequence view to identify which moments were in a fast food restaurant and which were not.

Our image sequence view showed its strength when there was one moment in which the lifelogger was having fast food in a restaurant. The concept appearing on the image was not a hamburger but a box of hamburger. Therefore, our detector was not able to detect this as a hamburger. Fortunately, our system can detect the concept "Chips" which is part of the "Main food" detector, we queried for this concept and find the image with that box of hamburger. Using the image sequence view, we searched back and forth from that moment and found the hamburger in the box (Figure 6). Finally, we reviewed our result images like previous topics (Figure 7).

Fig. 6: Moment where a hamburger is in the box and then gets unboxed

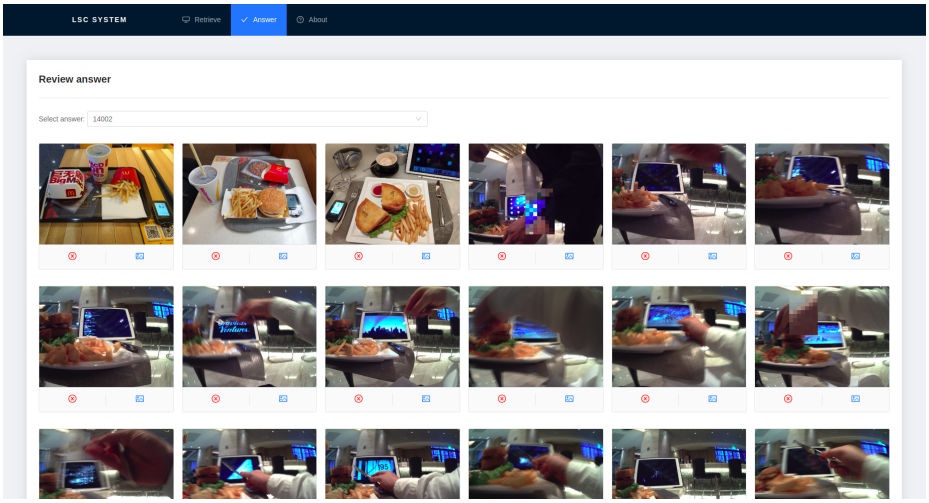

Fig. 7: Reviewing answer for topic: "Find the moment when u1 was eating fast food alone in a restaurant."

Find the moment when u1 was shopping for (and buying) a TV.

NOTE: *To be relevant, u1 must be shopping for a TV in an electronics store and actually buy the TV, later putting it in his car. Moments when u1 is simply looking at TVs without buying the TV are not considered relevant.*

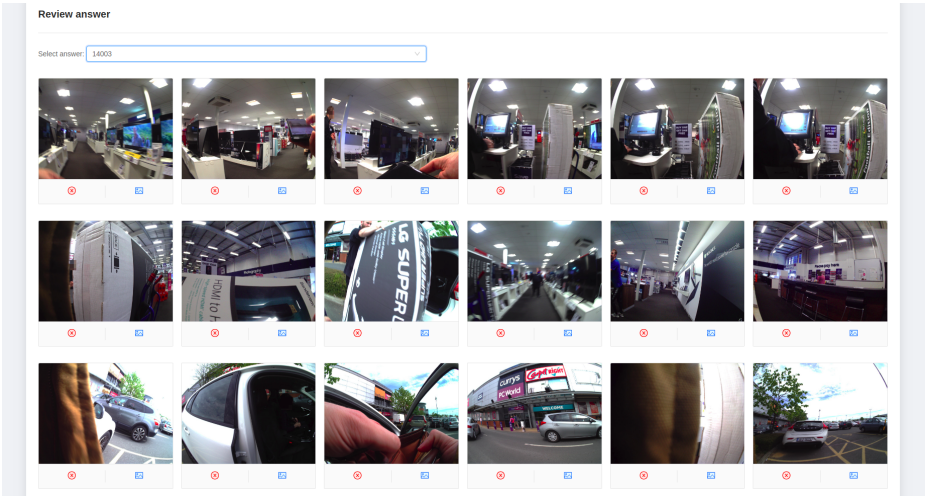

Fig. 8: Reviewing answer for topic: "Find the moment when u1 was shopping for (and buying) a TV."

This is one of the hard topics resolve because not only do we need to find moments with televisions in the lifelog dataset, we also have to make sure these moments are the correct ones when the description is very specific (the lifelogger buys the television, put in his car). Our COCO detector struggled to find the correct moments because it usually mistook a laptop for a television. Fortunately, our habit-based detector for concepts in "Technology" group can distinguish between a television and a laptop more efficiently and did a good job in detecting televisions in the lifelog dataset. This detection combining with image sequence view enabled us to easily find the correct moments (Figure 8).

Find the moment when u1 was looking at items in a toyshop.

NOTE: *To be considered relevant, u1 must be clearly in a toyshop. Various toys are being examined, such as electronic trains, model kits, and board games. Being in an electronics store, or a supermarket, are not considered to be relevant.*

The only relevant scene's category is toyshop, which is one of the categories in Place365 Standard dataset. Our system was able to retrieve all the moments when the lifelogger was shopping in a toyshop. There were two moments retrieved for this topic. Using image sequence view, we chose every relevant image which our system retrieved. We then finalized our result (Figure 9).

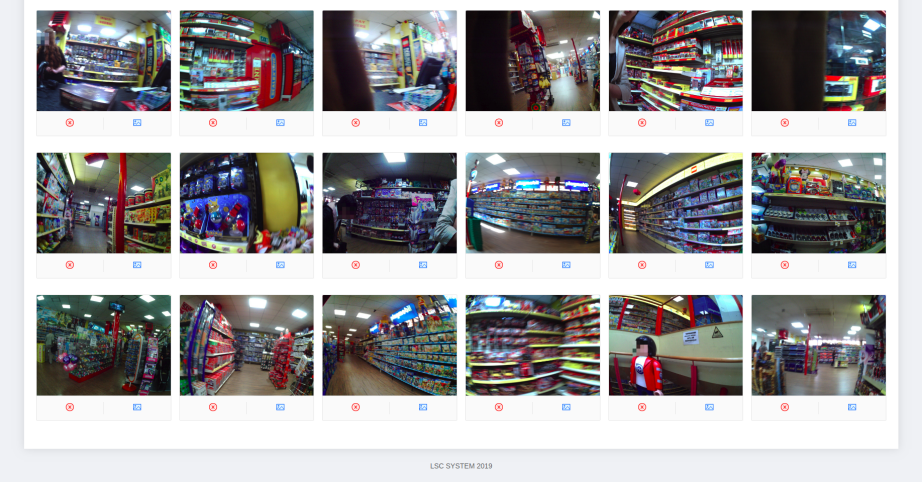

Fig. 9: Reviewing answer for topic: "Find the moment when u1 was looking at items in a toyshop."

Through the mentioned example queries, we aim to demonstrate the possible strategies for users to use our retrieval system in different scenarios. Depending on specific needs to query for a certain moment, a user can begin to retrieve related scenes based on scene's category, scene's attributes, or objects existing in images. Then the user can expand the sequence of images from a single image to further evaluate the context of the moment.

4.4 Result

In the LSAT subtask, participants were allowed to undertake the LAST task in an interactive or automatic manner. For interactive submissions, a maximum of five minutes of search time was allowed per topic. The LSAT task included 24 search tasks, generated by the lifeloggers and guided by Kahneman's lifestyle activities. We undertook this task in an interactive manner.

The result of our system's performance on the LSAT subtask is included in the NTCIR14 official result among other participants [5]. The official results are illustrated in the table below (Table 1). According to the result, our system has the highest performance among other runs in all metrics, including MAP, P@10, and RelRet.

Group ID	Run ID	Approach	MAP	P@10	RelRet
NTU	NTU-Run1	Interactive	0.0632	0.2375	293
NTU	NTU-Run2	Interactive	0.1108	0.3750	464
NTU	NTU-Run3	Interactive	0.1657	0.6833	407
DCU	DCU-Run1	Interactive	0.0724	0.1917	556
DCU	DCU-Run2	Interactive	0.1274	0.2292	1094
HCMUS	**HCMUS-Run1**	**Interactive**	**0.3993**	**0.7917**	**1444**
QUIK	QUIK-Run1	Automatic	0.0454	0.1958	232
QUIK	QUIK-Run2	Automatic	0.0454	0.1875	232

Table 1: The official results of the LSAT subtask

Although our system can achieve very promising results comparing to other systems, we still need to further improve our system to better represent unfamiliar concepts and integrate various interaction modalities to assist users in exploring lifelog data.

5 Conclusion

Our system supports the user to retrieve the moments of interest from the lifelog by 2 main steps: offline processing of the data (including annotating each image with metadata and structure the data for better performance and scalability), and optimize the user interaction (including a user-friendly design web application which supports flexible ways of searching and selecting results).

However, there are still some aspects that our system needs to improve. The user still needs to picture the moments to decide what scene category the images should be, and what concepts should be in the images.

In the future works, we will look into the aspect of natural language semantics to give our system the ability to understand the topic search and suggest more relevant inputs for the user.

Acknowledgements

This research is partially supported by research funding from Honors Program, University of Science, Vietnam National University - Ho Chi Minh City.

We would like to thank AIOZ Pte Ltd for supporting our research team with computing infrastructure.

References

1. LTA '16: Proceedings of the First Workshop on Lifelogging Tools and Applications. ACM, New York, NY, USA (2016)
2. LTA '17: Proceedings of the 2Nd Workshop on Lifelogging Tools and Applications. ACM, New York, NY, USA (2017)
3. Dinh, T.D., Nguyen, D., Tran, M.: Social relation trait discovery from visual lifelog data with facial multi-attribute framework. In: Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2018, Funchal, Madeira - Portugal, January 16-18, 2018. pp. 665–674 (2018). https://doi.org/10.5220/0006749206650674, https://doi.org/10.5220/0006749206650674
4. Dodge, M., K.R.: 'outlines of a world coming into existence': Pervasive computing and the ethics of forgetting. Environment and Planning B: Planning and Design 34(3), 431–445 (2007). https://doi.org/10.1068/b32041t
5. Gurrin, C., Joho, H., Hopfgartner, F., Dang-Nguyen, D.T., Zhou, L., Ninh, V.T., Le, T.K., Albatal, R., Healy, G.: Overview of the NTCIR-14 lifelog-3 task. In: Proceedings of the Fourteenth NTCIR conference (NTCIR-14) (2019)
6. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Albatal, R.: Overview of ntcir-12 lifelog task (2016)
7. Gurrin, C., Smeaton, A.F., Doherty, A.R.: Lifelogging: Personal big data. Found. Trends Inf. Retr. **8**(1), 1–125 (Jun 2014). https://doi.org/10.1561/1500000033, http://dx.doi.org/10.1561/1500000033
8. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Kamali, S., Malloci, M., Pont-Tuset, J., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://storage.googleapis.com/openimages/web/index.html (2017)
9. Le, N.K., Nguyen, D.H., Tran, M.T.: Smart lifelog retrieval system with habit-based concepts andmoment visualization. In: LSC 2019 @ ICMR 2019 (2019)
10. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
11. Nguyen, V.T., Le, K.D., Tran, M.T., Fjeld, M.: Nowandthen: A social network-based photo recommendation tool supporting reminiscence. In: Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia. pp. 159–168. MUM '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/3012709.3012738, http://doi.acm.org/10.1145/3012709.3012738

12. Tran, M., Truong, T., Duy, T.D., Vo-Ho, V., Luong, Q., Nguyen, V.: Lifelog moment retrieval with visual concept fusion and text-based query expansion. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. (2018), http://ceur-ws.org/Vol-2125/paper_109.pdf
13. Truong, T.D., Dinh-Duy, T., Nguyen, V.T., Tran, M.T.: Lifelogging retrieval based on semantic concepts fusion. In: Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge. pp. 24–29. LSC '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3210539.3210545, http://doi.acm.org/10.1145/3210539.3210545
14. Vo-Ho, V.K., Luong, Q.A., Nguyen, D.T., Tran, M.K., Tran, M.T.: Personal diary generation from wearable cameras with concept augmented image captioning and wide trail strategy. In: Proceedings of the Ninth International Symposium on Information and Communication Technology. pp. 367–374. SoICT 2018, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3287921.3287955, http://doi.acm.org/10.1145/3287921.3287955
15. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)