Lifelog Moment Retrieval with Advanced Semantic Extraction and Flexible Moment Visualization for Exploration

Nguyen-Khang Le¹, Dieu-Hien Nguyen¹, Vinh-Tiep Nguyen², and Minh-Triet Tran¹

 ¹ University of Science, VNU-HCM, Ho Chi Minh City, Vietnam {lnkhang, ndhien}@selab.hcmus.edu.vn, tmtriet@fit.hcmus.edu.vn
² University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam tiepnv@uit.edu.vn

Abstract. With the rise of technology over the last decade, the number of smart wearable devices, low-cost sensors as well as inexpensive data storage technologies have been increasing rapidly, making it easy for anyone to use them to capture the details of their everyday's life and create an enormous dataset which can include photos, videos, biometric and GPS information. These kinds of activities can be referred to as Lifelogging, which is becoming a popular trend in the research community. One of the most important tasks of processing lifelog data is to retrieve the moments of interest from the lifelog, which is also referred to as lifelog semantic access task. Our proposed system provides a novel way to extract semantic from the lifelog data using scene classification and object detection, including object color detection. In addition, we also design a user interface which can efficiently visualize moments in the lifelog. Using our solution, we achieve the first rank in Lifelog Moment Retrieval task of ImageCLEF Lifelog 2019 with F1@10 score of 0.61.

Keywords: Lifelog Retrieval \cdot Object Color Detection \cdot User Interaction.

1 Introduction

In a formal definition, a lifelog is the phenomenon whereby people record their own daily life in a varying amount of detail, for a variety of purposes. The record contains a more or less comprehensive dataset of a human's life and activities.

The concept of "Lifelog" has been around for a long time. A diarist named Robert Shields, who manually recorded 25 years of his life from 1972 to 1997 at 5-minute intervals in his diary, is probably the first ever lifelogger in the world.

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

He spent hours a day in the office recording data such as his body temperature, blood pressure, medications, .etc. In addition, he only slept for two hours at a time so he could be able to write about his dreams. His work consists of 37-million word diary which is also considered to be the longest ever written.

From an information science perspective, lifelogging provides us with huge archives of personal data. However, these are usually raw data which have no annotations, no semantic descriptions or even contain errors. Therefore, to make this data usable, it is a challenge to build a system that can understand the data semantically.

Nowadays, due to the dramatic increase of smart wearable devices. It is easier for 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.

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 [5] to retrieve the moments of interest from lifelog data can help people to revive memories [12], 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, whose mission is to retrieve moments of interest from the lifelog data.

We view this problem as two separate subproblems. The first subproblem aims to preprocess the data and annotate each data with appropriate metadata. We propose a way to extract basic and advanced concepts from the lifelog dataset. The second subproblem aims to design and provide a friendly user interface that enables novice users to interact with the queries and visualize the moment in a way that they can easily solve the search topic.

Comparing to our previous lifelog retrieval systems [10], the improvement on this system focus on more advanced concepts and more efficient user interaction. With this system, we efficiently solved the 10 test topics of ImageCLEF 2019 Lifelog (LMRT) and achieve the best result comparing to other runs.

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 [4]. 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.

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.

Lifelog is one of the four main tasks in ImageCLEF 2019 [7]. The main goal is to promote the evaluation of technologies for annotation, indexing and retrieval of visual data. The task aims to provide information access to large collections of images.

The annual Lifelog Search Challenge (LSC) takes more effort on the evaluation of interactive lifelog retrieval systems. In the LSC 2018 [1], associated with the dataset was 6 testing and 16 evaluation multimodal topics representing challenging real-world information needs.

In the previous version of our system [10], we proposed a retrieval system that is able to detect basic concepts. In our other work [9], we also conducted experiments on our system where a novice user uses the system to perform retrieval tasks.

Taking advantage of our previous works, we build a more efficient system that aims to detect more advanced concepts and improve the performance by applying modern techniques.

3 Proposed Retrieval System

3.1 Retrieval System Overview

To solve the retrieval 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 data processing problem and user interaction problem. We then process to solve these problems separately. The platform has two main processes (Figure 1):

1. The offline data processing aims to annotate each moment in the dataset with metadata that the user of the system can later use to retrieve the moment of interest. This process employs machine learning methods and vision-based algorithm to extract information about the visual scene, the concepts appearing on the image, etc., and use this information to annotate the image.



Fig. 1: Two main steps in our system: Offline Data Processing and Online Retrieval Process

2. The online retrieval process aims to provide a friendly user interface for the retrieval step which involves human interaction.

Our system architecture separate the back-end and the front-end, this will give us benefits in development, deployment and future improvement. For the back-end, we develop a RESTFul web service that performs retrieval tasks and provides data for its clients. The architecture of the back-end is based on the traditional multi-tier architecture. We employ three layer: the handler layer that handles requests and responses, the service layer that performs all the logic of our application, and data access layer which is in charge of reading data from and writing data to files and databases. The overview of the system architecture is illustrated in Figure 2

3.2 Retrieval System's Components

Main components of our system are illustrated in Figure 3. 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.



Fig. 2: Overview of our retrieval system architecture

For object detection, we employ a basic object detector trained on MS COCO 2014 dataset [11] and our habit-based detectors that take advantage of the Open Image V4 dataset [8]. Furthermore, a classifying model is trained to predict the scene's category and scene's attributes. In addition, we develop an object color detector to further improve our system.

3.3 Scene classification

To classify scenes in the lifelog dataset, we train the Residual Network (ResNet) on Places365-Standard dataset [13]. In this paper, ResNet152 has the highest top-5 accuracy when tested on the validation set and the test set compare to other three popular CNN architectures, AlexNet, GoogLeNet, and VGG 16 convolutional-layer CNN. With this model, we can annotate each image in the lifelog dataset with about 102 scene attributes and 365 scene categories.



Fig. 3: Main components of the proposed system (Lifelog retrieval system V2)

3.4 COCO object detector

To detect the concepts appearing on the images, we use the Faster R-CNN with a backbone of the 101-layer residual net (ResNet-101) and train our model on the MS COCO 2014 dataset (1) [11]. With this concept detector, we are able to detect 80 categories of concept. Furthermore, we group these categories into 11 super-categories as in the MS COCO [11].

3.5 Habit-based concept detector

We find that the lifeloggers, as well as any person, have several habits in their lives, they usually eat certain kinds of food, enjoy certain kinds of drink, and participate in certain activities. Because of this observation and our evaluation of the lifelog dataset, we prepare a number of detectors which aim to detect the concepts that appear many times in the daily life of the lifelogger in particular, and the people in his/her country in general.

3.6 Object color detection

Some search topic contains specific information about color of the object. Being able to detect an object with specific color in the lifelog dataset will improve our system performance a lot. Because of this reason, we develop our object color detector by applying Mask R-CNN [6] and K-means clustering. To detect the color of the object, we propose a method to find the object's dominant colors through clustering. We use "K-means Clustering" as the clustering technique. After applying the mask of the object on the image, we cluster pixels of the image by three channels: Red, Green, Blue. After the clustering process, the color at each cluster center will represent the color of that cluster, these colors will determine the dominant colors of the object.

Before applying K-means clustering method, we first standardize the variables (the values of red, green, blue channels) by dividing each data point by the standard deviation. This is for ensuring that the variations in each variable will affect the clusters equally.

We then perform clustering and experiment with several numbers of clusters to find the relevant number of clusters. We decide to have three clusters, representing three dominant colors of the object. Finally, we annotate the image in the dataset with this object colors.

Figure 4 demonstrates object detection and segmentation on the lifelog dataset. Taking the person wearing a black cloth and a blue jacket in Figure 4 for example, after applying the mask to the image, Figure 5 and Figure 6 illustrate the dominant colors of this object with 7 and 3 clusters respectively



Fig. 4: An example of detection and segmentation of objects in the lifelog dataset

3.7 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.



Fig. 5: Dominant colors extracted from the object using 7 clusters



Fig. 6: Dominant colors extracted from the object using 3 clusters

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.

Our retrieval system supports the user to automatically fill in input fields based on a description of the event in natural language. To build this feature, we first split the description into words. For each word, we find its position in the sentence (Noun, Verb, Adjective, Adverb, etc.), we only collect words which are nouns and verbs. After this step, we have a list of collected words, we then enhance this list by adding words that are in the thesaurus (synonyms and related concepts) of these words. The complete process is shown in Figure 7. A demonstration of this feature is shown in Figure 8.



Fig. 7: A complete process of extracting keywords from the event description to automatically fill in input fields

We employ the pagination strategy to visualize the resulting events to the user. This feature allows the user to visualize the result in pages, the user can navigate through each page and see the corresponding results. Also, the user can choose the number of results per page. The pagination process is conducted on the server side. Because the cost of resources for making a retrieval process is expensive, we do not execute this process whenever the user navigates through pages. Instead of that, the results of the retrieval are cached on the memory of the server. We use key-value caching, we use the information of the retrieval

Search >				
Search topic				
find the moment when the man was eating ice-cream outdoor, on a beau	h	G		
① Basic criteria				
Attributes :	Accuracy:			
outdoor ×	High v	?		
Categories :	Accuracy:			
beach × patio × coast ×	High v	?		
Concepts:				
man × ice_cream × outdoor × beach × cake × n	nirror ×	?		
open × field ×				
Search				

Fig. 8: A demonstration of the system's Automatic Input feature, keywords extracting from the search topic are automatically filled in

criteria (represented in a configured format) as keys and the retrieval results as values. Whenever the user executes a retrieval, the results are computed and cached. After that, when the user navigates through each page, we only perform the pagination on the cached data and return the results to the user. Finally, when the user makes the next retrieval with different criteria, we latest set of results on cache is cleared. Our cache keeps up to three sets of retrieval results. The complete flow is illustrated in Figure 9.



Fig. 9: A caching method to improve the performance of the system's pagination feature

4 Experiment with Queries in ImageCLEF Lifelog 2019

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 Task and Dataset

The detail of the dataset gathering process as well as the task's description is described in [2]. The task is split into two related subtasks using a completely new rich multimodal dataset which consists of 29 days of data from one lifelogger.

4.3 Search topic

Find the moment when user 1 was having breakfast at home.

NOTE: User 1 was having breakfast at home and the breakfast time must be from 5:00 AM until 9:00 AM.

Time is one of the most important aspects of this topic. The correct moments need to be in the range of time from 5:00 AM to 9:00 AM. Our system provides an easy way for the user to retrieve moments in an exact range of time in the day. Moreover, our food detector performs efficiently in the lifelog data and is able to detect moments when the lifelogger was having a meal. By inputting 5:00 AM, 9:00 AM as the time range, food as the concept appearing, and "Home" as the location's name, the user can retrieve every moment when the lifelogger was having breakfast at home in the morning.



Fig. 10: Result moments for topic: "Find the moment when user 1 was having breakfast at home."

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 11).



Fig. 11: 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.

Find the moment when u1 was having coffee with two person.

NOTE: Find the moment when u1 was having coffee with two person. One was wearing blue shirt and the other one was wearing white cloth. Gender is not relevant.

In this search topic, there are some advanced concepts such as blue shirt and white cloth, which indicate not only the objects but also their colors. Our system supports the user to search in the lifelog dataset for objects with basic colors. By search for concept "person" and the color "blue", the user of our system can efficiently retrieve the relevant moments (Figure 12).

4.4 Result in ImageCLEF Lifelog 2019

We participate in the ImageCLEF Lifelog 2019 - Lifelog Moment Retrieval task with the team name "HCMUS". In the ImageCLEF Lifelog 2019 [2], F1-



Fig. 12: Reviewing answer for topic: "Find the moment when u1 was having coffee with two person."

measure at 10 is used to evaluate the performance of participating systems. Figure 13 demonstrates the result of our team (HCMUS) in comparing with the results of other teams in Lifelog Moment Retrieval task of ImageCLEF 2019 Lifelog. According to the result, our system has the highest F1-measure at 10. The detail of our system performance is illustrated in Table 1, with Cluster Recall at 10 (CR@10), Precision at 10 (P@10), and F1-measure at 10 (F1@10)

Query	P@10	CR@10	F1@10
1	1	1	1
2	0.6	0.1	0.16
3	0.5	0.28	0.36
4	1	0.75	0.86
5	0.8	0.56	0.66
6	0	0	0
7	0.7	1	0.82
8	0.5	0.33	0.4
9	0.9	0.57	0.7
10	1	1	1

Table 1: Detail of our system performance in ImageCLEF Lifelog 2019 test topics

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.



Fig. 13: ImageCLEF 2019 Lifelog - LMRT Leaderboard. The result is evaluated by F1-measure at 10. Our system has the highest result comparing to other runs

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

We would like to thank AIOZ Pte Ltd for supporting our research team. This research is partially supported by the research funding for the Software Engineering Laboratory, University of Science, Vietnam National University -Ho Chi Minh City.

References

- 1. LSC '18: Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge. ACM, New York, NY, USA (2018)
- Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Zhou, L., Lux, M., Le, T.K., Ninh, V.T., Gurrin, C.: Overview of ImageCLEFlifelog 2019: Solve my life puzzle and Lifelog Moment Retrieval. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, vol. Vol-2380. CEUR-WS.org http://ceur-ws.org, Lugano, Switzerland (September 09-12 2019)
- 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)
- Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Albatal, R.: Overview of ntcir-12 lifelog task (2016)
- 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
- He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2980–2988 (2017)
- Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain,

J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)

- 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 largescale multi-label and multi-class image classification. Dataset available from https://storage.googleapis.com/openimages/web/index.html (2017)
- Le, N.K., Nguyen, D.H., Hoang, T.H., Nguyen, T.A., Truong, T.D., Dinh, D.T., Luong, Q.A., Vo-Ho, V.K., Nguyen, V.T., Tran, M.T.: Hcmus at the ntcir-14 lifelog-3 task. In: 14th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-14 2019). Tokyo, Japan (June 10-13 2019)
- Le, N.K., Nguyen, D.H., Hoang, T.H., Nguyen, T.A., Truong, T.D., Dinh, D.T., Luong, Q.A., Vo-Ho, V.K., Nguyen, V.T., Tran, M.T.: Smart lifelog retrieval system with habit-based concepts and moment visualization. In: LSC 2019 @ ICMR 2019. Ottawa ON, Canada (June 10 - 13 2019)
- 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)
- Nguyen, V.T., Le, K.D., Tran, M.T., Fjeld, M.: Nowandthen: A social networkbased 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)
- 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)