Organizer Team at ImageCLEFlifelog 2017: Baseline Approaches for Lifelog Retrieval and Summarization

Liting Zhou[1], Luca Piras[2], Michael Riegler[3], Giulia Boato[4], Duc-Tien Dang-Nguyen[1], and Cathal Gurrin[1]

[1] Insight Centre for Data Analytics, Dublin City University
zhou.liting2@mail.dcu.ie, {duc-tien.dang-nguyen, cathal.gurrin}@dcu.ie
[2] DIEE, University of Cagliari
luca.piras@diee.unica.it
[3] Simula Research Laboratory
michael@simula.no
[4] DISI, University of Trento
boato@disi.unitn.it

Abstract. This paper describes the participation of Organizer Team in the ImageCLEFlifelog 2017 Retrieval and Summarization subtasks. In this paper, we propose some baseline approaches, using only the provided information, which require different involvement levels from the users. With these baselines we target at providing references for other approaches that aim to solve the problems of lifelog retrieval and summarization.

1 Introduction

Personalized multimedia archives that contain a large amount of data collected using various personal devices, such as smart phones, cameras, wearable devices and so on are getting more and more common nowadays. In these archives, every moment and aspect of our lives are stored. They can contain information about our daily routines, consumed food but also about our health status, etc. These data logs of a human lives, also commonly referred to as lifelogs, are more and more interesting for the research community but also companies. Collecting and storing the data is one challenge but getting insights from the collected data and find new information by connecting different types of data requires a lot of researches for analyzing, categorizing and querying these huge amounts of data in a efficient way.

In this paper we present our approach to tackle the Image CLEF 2017 [11] Lifelog Task [6], which aims at solving the problems of lifelog retrieval and summarization. Lifelogs are usually chronologically organized and moments that belong to the same activity or the same event are normally very similar. This can be exploited to reduce processing time by grouping moments that are similar based on the time when they happened and the belonging concepts. This transforms the image retrieval challenge into a image segments retrieval challenge.

This has the advantage that boundaries between moments or activities are automatically segmented based on time and concepts [7]. To remove non-relevant images filtering is recommended. In our case, we remove images that seem to be sparse on information (blurry, only big objects, etc.) Retrieved images then can be diversified into clusters which then can be further used for summarization, which can be done automatically or via relevance feedback by follow the methods described in [5].

The remainder of this paper is organized as follows, first we present related work in the field. This is followed by a detailed description of our approach. After that we present the experimental results which is followed by a discussion and conclusion.

2 Related Work

In this section we discuss briefly recent studies on lifelog segmentation and the retrieval problem in term of relevant and diversity. In addition, many novel techniques are proposed and evaluated, to accurately retrieve the similar events from lifelog dataset using contextual data.

Typically, chronological images segmentation is done by heuristic split based on long interval with no capture [14] or by thresholding the distances between the frames (or images) based on the content [3]. Doherty et al. in [8], to determine the similarity between adjacent block of images, proposed to use Hearst's Text Tiling Algorithm [9] on edge histogram, which is extracted by using Canny edge detection. For egocentric photo streams (from wearable cameras), a typical segmentation is based on unsupervised hierarchical agglomerative clustering to extract the key-frame summary [1].

Current works in multimedia retrieval have considered relevance and diversity as two core criteria. Relevance was commonly estimated based on textual information, e.g., from the photo tags, and many of current search engines are still mainly based on this information. Diversity is usually improved by applying clustering algorithms which rely on textual or/and visual properties [12]. Recently, in social image retrieval, some methods have exploited the participation of humans by collecting the feedbacks of the results to improve the diversification [2]. To reduce the number of images to be returned to the user, some papers in the past years proposed the use of image clustering techniques [15]. These approaches exploit the hierarchical indexing structure of the clusters to refine the number of images to consider [13]. More recently different type of relevance feedback has been used to expand the query to improve both relevant and diversity as well as reduce the number of iterations [5].

3 The proposed approaches

The proposed approaches follow the schema as illustrated in Figure 1. Since lifelogs are chronologically organized and moments in the same activity or the same

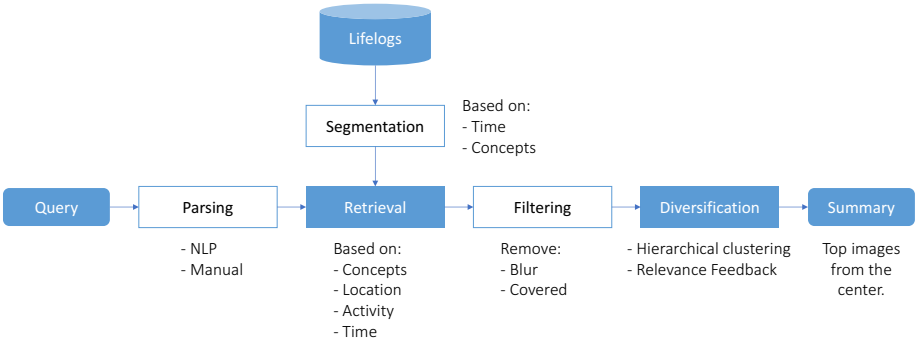

Fig. 1. Schema of the proposed methods.

event are normally very similar to each other, in order to reduce the processing time, we group similar moments together based on time and concepts. By applying this chronological-based segmentation, we turn the problem of images retrieval into image segments retrieval, in which the boundary between activities such as having breakfast, working in front of a computer, and so on [7], are automatically decided based on the time and concepts. Starting from a topic query, it is transformed into small inquiries, where each of them is asking for a single piece of information of concepts, location, activity, and time. The moments that matched all of those requirements are returned as the retrieval results. In order to remove the non-relevant images, a filtering step is applied on the retrieved images, by removing blurred and images that covered mainly by huge object or by the arms of the user. Finally, the images are diversified into clusters and the top images that close to center are selected for the summarization, which can be done automatically or using relevance feedback by follow the methods in [5]. These steps are described as follows:

3.1 Segmentation

For the segmentation we applied a simple chronological-based segmentation as follow: For each pair of two continuous images I_t and I_{t+1} at the time t, the distance $d(I_t, I_{t+1})$ between them is computed as:

$$d(I_t, I_{t+1}) = ||\boldsymbol{C_t} - \boldsymbol{C_{t+1}}||$$

where $||\cdot||$ is the normalized Euclidean distance, and $\boldsymbol{C}$ is the concept vector of each image provided from the task. If $d(I_t, I_{t+1}) < \tau$, where τ is a threshold, the two images are set belong to the same segment, otherwise they are set in different segments. If τ is too small, an activity should be split into small activities, while larger value of tau should grouped different activities into the same one. Since $||\cdot||$ is normalized, when $\tau = 0$, the images are grouped into different segments, and when $\tau = 1$, all images are belong to a single segment.

Segmenting the activities is not simply an incident of identifying the exact event boundaries; it also concerns with keeping track of the fine-grained group of events together into extended meaningful units, and thus deciding the right value of τ is not trivial. In the proposed approaches, we try different values of *tau* for different runs, which will be explained in Section 4.

After this step, each segment is represented by the first image (of that segment) with these basic information: location, activity, time segment, number of people and the list of the concepts. If any of these information is missing from the first image, we take it from the second image and so on.

3.2 Parsing the Query and Retrieval

Converting a topic into precise criteria for retrieval is the key question for both sub-tasks. It can be automatically done by considering any word in the topic as the queried concepts and then searching for all segments that contain those concepts; by applying natural language processing techniques; or by fine-tunning by a human in the loop, i.e., the user will read the topic and "translate" it into the search criteria. For example, with the topic:

- `Topic: Using laptop out of office`
- `Query: Find the moment(s) in which user u1 was using his laptop outside the working places.`
- `Description: To be consider to relevant, the user should use his laptop, for work or for entertainment out of his working place.`

Can be "translated into":

- `User: +u1`
- `Concepts: +laptop`
- `Activities: +working`
- `Time: --`
- `Location: -work`

where `+/-` means the retrieval images has to contain/not contain this information, respectively, and `--` means any.

In the proposed approaches, we use the automatic and the human-in-the-loop methods. Our "translation" will be shown in Section 4.

3.3 Filtering

An image can be considered as blurred based on it focus level. In the proposed approaches, we estimate the focus by computing the absolute sum of the wavelet coefficients and comparing it to a threshold, by exploiting the method in [10]. The return of this method is a scalar number in $[0, 99]$ which the bigger value the sharper image. From our observation, for values below 30, most of the images are blurred, and thus we set this threshold to 30.

In order to remove images that covered by large objects, we apply an heuristic method as follows:

Step 1 Convert the image to binary images by applying thresholding with several thresholds.
Step 2 Extract connected components and calculate their centers.
Step 3 Group centers based on their coordinates, and then close them to form the corresponds blob.
Step 4 Take the biggest blob and its size (in pixels).

If the size is over 50% of the whole area, the image is considered as covered. This whole method is implemented by calling the function `SimpleBlobDetector` from OpenCV[5].

After this step, all remain images are considered as relevant to the topic. Please notice that the images are still kept inside the segment.

3.4 Diversification

In this step, for automatic approach, we use a hierarchical agglomerative clustering algorithm (see in [4]) to group similar segments into the same cluster based on the concepts. The clusters are then sorted based on the number of segments, decreasingly. Finally, we produce the summary for the queried by selecting representative images from the clusters based by selecting the images closest to the center of each cluster.

We also propose a human-in-the-loop approach in this step by using the usual dichotomous Relevance Feedback paradigm (more details can be seen in [5]), that asks the user to assign the labels *Relevant* \ *Non-relevant* to the retrieved images. The system asks the user to label the representative images of the top N results returned by the automatic diversification procedure (as mentioned above), and the number of images that have been labeled as being *Relevant* \ *Non-relevant* for each cluster is computed. Then, the clusters are sorted as follows:

- Clusters that have a large number of relevant counts are sorted higher.
- Clusters that have the same number of relevant counts are sorted based on the number of non-relevant counts (i.e., a cluster that contains a larger number of 'non-relevant' images should be selected later).
- Clusters that have the same number of *Relevant* \ *Non-relevant* counts are sorted on the basis of the number of segments.

For each cluster, the images that are selected to represent the topic are chosen in the same way as in the automatic diversification.

4 Experimental results

4.1 Submitted Runs

We submitted 3 runs on the Retrieval sub-task and 5 runs on the Summarization sub-task, summarized in Table 1.

As for the retrieval task, the first run is exploiting only time and the concepts information. We consider every single image as the basic unit and the retrieval just returns all images that contains the concepts extracted from the topics. We named this run is the 'baseline' with the purpose that any other approaches should obtain better performance than this.

Table 1. Submitted Runs.

RunID	Name	τ	Parsing	Filtering	Diversification
LRT Run 1	Baseline	0	Automatic	–	–
LRT Run 2	Segmentation	0.05	Automatic	–	–
LRT Run 3	Fine-tunning	0.05	Fine-tunning	–	–
LST Run 1	Baseline	0	Automatic	Not apply	Automatic
LST Run 2	Segmentation	0.05	Automatic	Not apply	Automatic
LST Run 3	Filtering	0.05	Automatic	Apply	Automatic
LST Run 4	Fine-tunning	0.05	Fine-tunning	Apply	Automatic
LST Run 5	Relevance Feedback	0.05	Fine-tunning	Apply	Relevance Feedback

With the second run, we applied the optimized value for τ (optimized from the devset) to do the segmentation. So in this run, the only difference is the basic unit of retrieval now is the segment, not image.

For the Fine-tunning runs, the "translation" is applied as in Tables2, and 3.

Table 2: Parsing as a fine-tuning on testset, LRT subtask. + means selection and – means exception.

Topic	User	Activities	Times	Locations	Concepts
T001	u1	-Walking, -Running	+MinuteID: 400-1400	-Work, +Home, +Science Gallery Caf, + Helix	+Laptop
T002	u1	-Walking, -Running, -Transport	+MinuteID: 720-1080(workday)	-Work, -Home	+Microphone
T003	u1	-Walking, -Running, -Transport	+MinuteID: 540-1080	-Work, -Home, +Dublin Airport (DUB)	+Hard disc, +Knee pad, +Mouse, +CD player
T004	u1	+Running, -Walking	+MinuteID: 400-660(weekend)	-Work, -Home, +Place in Saint Anne's Park, +Hampstead Park	+Park bench
T005	u1	-Walking, -Running, -Transport	+MinuteID: 540-1140(workday)	+Work, -Home	+Table, +Laptop
T006	u1	-Walking, -Running, -Transport	+MinuteID: 400-540(workday), 1140-1400(workday), 400-1400(weekend)	-Work, +Home	+TV

[5] http://opencv.org

T007	u1	-Walking, -Running, -Transport	+MinuteID: 400-540, 660-840, 1080-1190	-Home, -Work, +Science Gallery Caf, +DCU Restaurant,	-Television, -laptop, -Commic book, -Notebook
T008	u1	+Transport, -Walking, -Running	+MinuteID: 400-1400	-Work, -Home	+Laptop
T009	u2	-Walking, -Running, -Transport	+MinuteID: 590-1400(workday), 540-1240(weekend)	-Work, +Home	+Guitar
T010	u2	-Transport, +Running, +Walking	+MinuteID: 960-1190	+DCU, -Home, -Work	+Running shoes
T011	u2	-Waking, -Running, -Transport	+MinuteID: 540 -1240	+Work, +Home	Apple, +Banana, + Orange, + Strawberry
T012	u2	-Waking, -Running, -Transport	+MinuteID: 400-540, 1290-1400	+Home, -Work	+Washbasin
T013	u2	-Waking, -Running, -Transport	+MinuteID: 400-1400	+DCU, +Home, +Starbucks, +Costa Coffee	+Banana, +Apple, +Peach, +Broccoli, +Spaghetti squash, +Cheeseburger, +Hotdog, + Mashed potato
T014	u2	-Waking, -Running, -Transport	+MinuteID: 400-540, 660-840, 1080-1190	+McDonald's	+Cheeseburger
T015	u2	+Walking, -Running, -Transport	+MinuteID: 540-1080	-Work, -Home, +Place in Yong He Gong Lama Temple, +Place in Confucian Temple	+N/A
T016	u2	-Walking, -Running, -Transport	+MinuteID: 540-1138	-Work, -Home	+ATM
T017	u3	-Walking, -Running, -Transport	+MinuteID: 400-1400	-Work, -Home	+Wine bottle, +Beer bottle, +Beer glass
T018	u3	+Walking, -Running, -Transport	+MinuteID: 650-1080	+ Lidl, +Butchery	+butcher shop, meat market

T019	u3	-Walking, -Running, -Transport	+MinuteID: 540-1140	-Work, -Home	+Vending machine
T020	u3	+Walking, -Running, -Transport	+MinuteID: 590 - 1140	-Work, -Home, +Lidl, +Butchery, +Place in Dublin 1	+Butcher shop, +CD player, +Shoe shop, +Toyshop, +Bakeshop, +Grocery store

Table 3: Parsing as a fine-tuning on testset, LST subtask.
+ means selection and – means exception.

Topic	User	Activities	Times	Locations	Concepts
T001	u1	-Walking, -Running, -Transport	+MinuteID: 540-1140(workday)	+Work, -Home	+Table, +Laptop
T002	u1	-Walking, -Running, -Transport	+MinuteID: 400-540(workday), 1140-1400(workday), 400-1400(weekend)	-Work, +Home	+TV
T003	u1	-Walking, -Running	+MinuteID: 400-1400	-Work, +Home, +Science Gallery Caf, + Helix	+Laptop
T004	u1	-Walking, -Running, -Transport	+MinuteID: 400-540, 1140-1400	+Home	+Laptop, +Notebook
T005	u2	-Waking, -Running, -Transport	+MinuteID: 400-1400	+DCU, +Home, +Starbucks, +Costa Coffee	+Banana, +Apple, +Peach, +Broccoli, +Spaghetti squash, +Cheeseburger, +Hotdog, + Mashed potato
T006	u2	-Waking, -Running, -Transport	+MinuteID: 720-1400(weekend), 960-1400(workday)	-Work,-Home	+Wine bottle, +Beer bottle, +Beer glass
T007	u2	-Transport, +Walking, -Running	+MinuteID: 400-1140	-Work, -Home, +Place in Beijing, +Place in Yong He Gong Lama Temple, +Place in Chaoyang	N/A

T008	u2	+Transport, -Walking, -Running	+MinuteID: 430-590, 1080-1190	N/A	N/A
T009	u3	-Transport, -Walking, -Running	+MinuteID: 400-540, 660-840, 1080-1190	+Home, -Work	+Frying pan, +Pot
T010	u3	-Transport, +Walking, -Running	+MinuteID: 590 - 1140	-Work, -Home, +Lidl, +Butchery, +Place in Dublin 1	+Butcher shop, +CD player, +Shoe shop, +Toyshop, +Bakeshop, +Grocery store

The same strategy is applied on the summarization subtask, in which the first three runs were ran to test the automatic approach with the increasing level of the 'criteria', while the last two runs are used to test the fine tuning and the relevance feedback approaches. For the relevance feedback approach, we ran a simulation by exploiting the ground-truth annotated data.

4.2 Results

Shown in Tables 4 and 5 are the results of the runs on the retrieval and summarization sub-tasks, respectively. The results confirm that applying segmentation improved both retrieval and summarization performance. It is quite clear that applying fine-tunning significantly improved the performance. The big gaps in results between the automatic approach with the fine-tunning and between the fine-tunning with the human-in-the-loop (relevance feedback) approaches, shown that we need better natural language processing as well as machine learning studies for these problems.

Table 4. Lifelog Retrieval Results.

Run	Name	Average NDCG
LRT Run 1	Baseline	0.09
LRT Run 2	Segmentation	0.14
LRT Run 3	Fine Tuning	0.39

5 Discussions and Conclusions

In this paper we introduced different baseline approaches, that came from fully automatic to fully manual paradigm, proposed by the Organizer Team of the ImageCLE-Flifelog 2017 task as participant of the Retrieval and Summarization subtasks. These approaches, that require different level of involvement of the users, exploit only the information provided by the organizers along with the collection of images, i.e., the description of the semantic locations and the physical activities. From the obtained results it appears clear that deeper analysis of the methods should be considered as well as the use of extra information.

Table 5. Lifelog Summarization Results.

Run	Name	Average F1@10
LST Run 1	Baseline	0.10
LST Run 2	Segmentation	0.17
LST Run 3	Filtering	0.18
LST Run 4	Fine Tuning	0.32
LST Run 5	Relevance Feedback	0.77

References

1. Bolanos, M., Mestre, R., Talavera, E., Nieto, X.G., Radeva, P.: Senseseer mobile-cloud-based lifelogging framework. Visual summary of egocentric photostreams by representative keyframes pp. 1–6 (July 2015)
2. Boteanu, B., Mironic, I., Ionescu, B.: A relevance feedback perspective to image search result diversification. In: 2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP). pp. 47–54 (Sept 2014)
3. Byrne, D., Lavelle, B., Doherty, A.R., Jones, G.J., Smeaton, A.F.: Using bluetooth and gps metadata to measure event similarity in sensecam images. 5th International Conference on Intelligent Multimedia and Ambient Intelligence (July 2007)
4. Dang-Nguyen, D.T., Piras, L., Giacinto, G., Boato, G., De Natale, F.G.: A hybrid approach for retrieving diverse social images of landmarks. In: 2015 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2015)
5. Dang-Nguyen, D.T., Piras, L., Giacinto, G., Boato, G., De Natale, F.G.: Multimodal retrieval with diversification and relevance feedback for tourist attraction images. ACM Transactions on Multimedia Computing, Communications, and Applications (2017), accepted
6. Dang-Nguyen, D.T., Piras, L., Riegler, M., Boato, G., Zhou, L., Gurrin, C.: Overview of ImageCLEFlifelog 2017: Lifelog Retrieval and Summarization. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
7. Doherty, A.R., Smeaton, A.F.: Automatically segmenting lifelog data into events. 9th International Workshop on Image Analysis for Multimedia Interactive Services (30 June 2008), http://doras.dcu.ie/4651/
8. Doherty, A.R., Smeaton, A.F., Lee, K., Ellis, D.P.: Multimodal segmentation of lifelog data. Large Scale Semantic Access to Content (Text, Image, Video, and Sound) pp. 21–38 (June 2007), http://dl.acm.org/citation.cfm?id=1931393
9. Hearst, M.A.: Texttiling: A quantitative approach to discourse segmentation. Technical Report UCB:S2K-93-24 (1993)
10. Huang, J.T., Shen, C.H., Phoong, S.M., Chen, H.: Robust measure of image focus in the wavelet domain. In: Intelligent Signal Processing and Communication Systems. pp. 157–160 (2005)
11. Ionescu, B., Müller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D.T., Dicente Cid, Y., Eickhoff, C., Garcia Seco de Herrera, A., Gurrin, C., Islam, B., Kovalev, V., Liauchuk, V., Mothe, J., Piras, L., Riegler, M., Schwall, I.: Overview of ImageCLEF 2017: Information extraction from images. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017. Lecture Notes in Computer Science, vol. 10456. Springer, Dublin, Ireland (September 11-14 2017)

12. van Leuken, R.H., Garcia, L., Olivares, X., van Zwol, R.: Visual diversification of image search results. In: Proceedings of the 18th International Conference on World Wide Web. pp. 341–350. WWW '09, ACM, New York, NY, USA (2009)
13. Mironica, I., Ionescu, B., Vertan, C.: Hierarchical clustering relevance feedback for content-based image retrieval. In: IEEE International Workshop on Content-Based Multimedia Indexing. pp. 1–6 (2012)
14. Peitgen, H., Jürgens, H., Saupe, D.: Chaos and fractals - new frontiers of science (2. ed.). Springer (2004)
15. Thomee, B., Lew, M.S.: Interactive search in image retrieval: a survey. International Journal of Multimedia Information Retrieval 1(1), 71–86 (2012)