2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)

An Interactive Watershed-based Approach for Lifelog Moment Retrieval

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

Minh-Son Dao, Koji Zettsu
Big Data Analytics Laboratory
National Institute of Information and Communications Technology
Tokyo, Japan
{*dao, zettsu*} @*nict.go.jp*

***Abstract*—Recently, terminologies "lifelogging" and "lifelog" became to represent the activity of continuously recording people everyday experiences and dataset contained these recorded experiences, respectively. Hence, providing an excellent tool to retrieve life moments from lifelogs to fast and accurately bring a memory back to a human when required, become a challenging but exciting task for researchers. In this paper, a new method to meet this challenge by utilizing the hypothesis that a sequence of images taken during a specific period can share the same context and content is introduced. This hypothesis can be explained in another way that if there is one image satisfies a given query (i.e., seed); then a certain number of its spatiotemporal neighbors probably can share the same content and context (i.e., watershed). Hence, an interactive watershed-based approach is applied to build the proposed method that is evaluated on the imageCLEFlifelog 2019 dataset and compared to participants joined this event. The experimental results confirm the high productivity of the proposed method in both stable and accuracy aspects as well as the advantage of having an interactive schema to push the accuracy when there is a conflict between a query and how to interpret such a query.**

***Index Terms*—lifelogs, watershed, clustering, image retrieval, feature extracting, content and context, semantic**

I. Introduction

One of the most exciting topics when trying to get insights from lifelogs is to understand human activities from the first-person perspective [1] [2]. Another interesting topic is to augment human memory towards improving human capacity to remember [3]. The former aims to understand how people act daily towards having effective and efficient support to improve the qualification of living both in social and physical activities. The latter tries to create a memory assistant that can accurately and quickly bring a memory back to a human when necessary. Both of these topics need fast and accurate lifelog moments retrieval (LMTR) systems.

To create a competitive and collaborative environment for encouraging research focused on LMTR two annual events have established: (1) NTCIR-LMTR, and (2) imageCLEFlifelog. The former focusing on online LMTR when participants have a limited time to answer given queries quickly. The latter gives participants several months to develop new and exciting solutions.

Among participants joined these events, AILab-GTI [4] and RegimLab [5] both utilized AlexNet [6] and GoogleNet [7] to extract visual features. VCI2R [8] has used a wide range of CNN-based detector and classifier to collect visual features. Their methods aim to build a classifier for events by getting deep-learning-based visual features and retraining on their data. Not satisfied with using only visual features, metadata are integrated with visual features to increase the accuracy of retrieving. Representing for this direction is Zhou et al. [9], NLP-Lab [10], Regim Lab [5], and HCMUS [11].

In general, visual concepts are popular among methods above where visual features are converted to visual concepts to meet a querys content. Nevertheless, the qualification of these visual concepts depends totally on used detectors. For example, a detector trained on COCO dataset [12] has only 80 object categories probably is not enough to describe all objects in lifelog data.

In this paper, we introduce a new method aimed to tackle LMTR challenge. The proposed method is evaluated on and compared to other participants of the imageCLEFlifelog2019 challenge. The proposed method is built based on the hypothesis that a sequence of images taken during a specific period can share the same context and content. The interactive schema is also introduced as a useful tool to overcome the automatic misinterpretation from given queries.

II. Methodology

In this section, we introduce the proposed method, as well as its detail explanation, algorithms, and examples.

A. The Proposed Method

The proposed method is built based on the following observation: A sequence of images taken during a specific period can share the same context and content. Thus, if a set of images can be clustered into sequential atomic clusters, given an image, we can automatically find all images sharing the same content and context by first finding the atomic cluster sharing the same content, then watershed reward and forward to find other clusters sharing the same context. The terminology atomic cluster is understood that all images inside should share the similarity higher than the predefined threshold and must share the same context (e.g., location, time).

978-1-7281-5527-2/19/$31.00 ©2019 IEEE
DOI 10.1109/BigMM.2019.00052

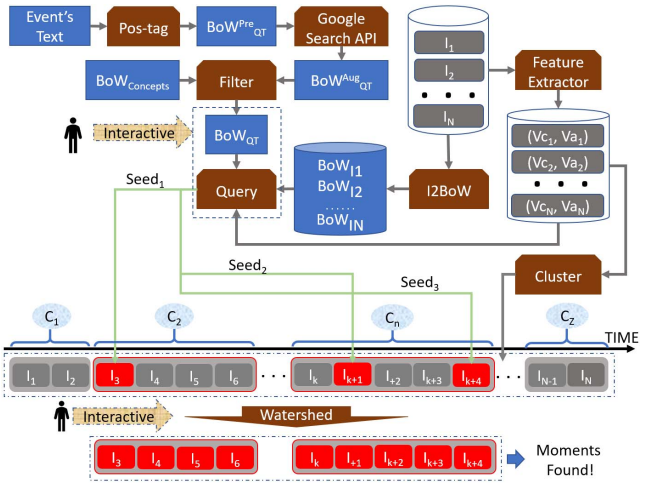

Fig. 1. An Overview of the Proposed Method

Based on the above discussion, the Algorithm 1 is built to find a Lifelog moment from a dataset defined be a query events text. Following is the description to explain the Algorithm 1.

- **Stage 1** (Offline): This stage aims to cluster a dataset into the atomic clusters. The category vector Vc and attribute vector Va extracted from images are utilized to build the similarity function. Besides, each image is analyzed by using *I2BoW* function to extract its BoW contains concepts and attributes reserved for later processing.
- **Stage 2** (Online): This stage targets to find all clusters satisfied a given query. Since the given query is described by text, a text-based query method must be used for finding related images. Thus, we create *Pos_tag*, *GoogleSearchAPI*, and *Filter* functions to find the best BoWs that represent the taxonomy of the querys context and content. In order to prune the output of these functions, we utilize the *Interactive* function to select the best taxonomy. First, images queried by using BoW (i.e., seeds) are utilized for finding all clusters, namely LMTR1, that contain these images. These clusters are then hidden for the next step. Next, these seeds are used to query on the rest clusters (i.e., unhidden clusters) to find the second set of seeds. These second set of seeds are used to find all clusters, namely LMTR2, that contain these seeds. The final output is the union of C1 and C2. At this step, we might apply *Interactive* function to refine the output (e.g., select clusters manually). Consequently, all lifelog moment satisfied the query are found.

B. Functions

In this subsection, the significant functions utilized in the proposed method are introduced, as follows:

- **Pos_tag**: processes a sequence of words tokenized from a given set of sentences, and attaches a part of speech tag (e.g., noun, verb, adjective, adverb) to each word. The library NLTK[1] is utilized to build this function.
- **featureExtractor**: analyzes an image and return a pair of vectors v_c (category vector, 512 dimension) and v_a (attribute vector, 102 dimension). The former is extracted from the "avgpool" layer of the RestNet18 of the pre-trained model PlaceCNN [13]. The latter is calculated by using the equation $v_a = W_a^T v_c$, introduced in [14], where W_a is the weight of Learnable Transformation Matrix.
- **similarity**: measures the similarity between two vectors using the cosine similarity.
- **I2BoW**: converts an image into a bag of words using the method introduced in [15]. The detector developed by the authors return concepts, attribute, and relation vocab. Nevertheless, only a pair of attribute and concept is used for building I2BoW.
- **GoogleSearchAPI**: enriches a given set of words by using Google Search API [2]. The output of this function is the set of words that could be probably similar to the queried words under certain concepts.
- **Interactive**: allows users to interfere with refining the results generated by related functions.
- **Query**: finds all items of the searching dataset that are similar to queried items. This function can adjust its similarity function depending on the type of input data.
- **cluster**: clusters images into sequential clusters so that all images of one cluster must share the highest similarity comparing to its neighbors. All images are sorted by time before being clustered. Algorithm 2 describes this function in detail. Figure 1 illustrates how this function works.

C. Parameter Definitions

Let $I = \{I_i\}_{i=1..N}$ denote the set of given images (e.g., dataset).

Let $F = \{(Vc_i, Va_i)\}_{i=1..N}$ denote the set of feature vectors extracted from I.

Let $C = \{C_k\}$ denote a set of atomic clusters.

Let $\{BoW_{I_i}\}_{i=1..N}$ denote the set of BoWs; each of them is a BoW built by using the *I2BoW* function.

Let $\{BoW_{QT}\}, \{BoW_{QT}^{Noun}\}, \{BoW_{QT}^{Aug}\}$ denote the Bag of Words extracted from the query, the NOUN part of BoW_{QT}, and the augmented part of BoW_{QT}, respectively.

Let $Seed_j^i$ and $LMTR^k$ denote a set of seeds and lifelog moments, respectively.

III. Experimental Results

In this section, the archive (datasets and queries), and evaluation metrics used to evaluate the proposed solution are introduced. Besides, the comparison of the proposed method with others is also discussed.

A. Archive and Evaluation metrics

We use the dataset released by the imageCLEFlifelog 2019 challenge [16]. The challenge introduces an entirely new rich multimodal dataset which consists of 29 days of data from

[1]https://www.nltk.org/book/ch05.html

[2]https://github.com/abenassi/Google-Search-API

Algorithm 1 A Interactive Watershed-based Lifelog Moment Retrieval

Input: QT, $BoW_{Concepts}$, $\{I_i\}_{i=1..N}$
Output: $LMTR$

{OFFLINE}
$\{BoW_{I_i}\}_{i=1..N} \Leftarrow \emptyset$
$\{(Vc_i, Va_i)\}_{i=1..N} \Leftarrow \emptyset$
$\forall i \in [1..N], BoW_{I_i} \Leftarrow I2BoW(I_i)$
$\forall i \in [1..N], (Vc_i, Va_i) \Leftarrow featureExtractor(I_i)$
$\{C_m\} \Leftarrow cluster(\{I_i\}_{i=1..N})$ using $\{(Vc_i, Va_i)\}$
{ONLINE}
$BoW_{QT}^{Noun} \Leftarrow Pos\_tag(QT).Noun$
$BoW_{QT}^{Aug} \Leftarrow GoogleSearchAPI(BoW_{QT}^{Noun})$
$BoW_{QT}^{Noun} \Leftarrow Filter(BoW_{QT}^{Aug} \cap BoW_{Concepts})$
$BoW_{QT} \Leftarrow Interactive(BoW_{QT}^{Noun}, Pos\_tag(QT))$
$\left\{Seed_j^1\right\} \Leftarrow Interactive(Query(BoW_{QT}, \{BoW_{I_i}\}_{i=1..N}))$
$LMTR^1 \Leftarrow \left\{C_k | \forall j \in \left\|\left\{Seed_j^1\right\}\right\|, Seed_j^1 \in \{C_k\}\right\}$
$\left\{C_l^{rem}\right\} \Leftarrow \{C_m\} - LMTR^1$
$\left\{Seed_j^2\right\} \Leftarrow Query(\left\{Seed_j^1\right\}, \{C_l^{rem}\})$
$LMTR^2 \Leftarrow \left\{C_l | \forall j \in \left\|\left\{Seed_j^2\right\}\right\|, Seed_j^2 \in \{C_l^{rem}\}\right\}$
$LMTR \Leftarrow LMTR^1 \cup LMTR^2$
$LMTR \Leftarrow Interactive(LMTR)$
return $LMTR$

one lifeloggers. The dataset contains images (1,500-2,500 per day from wearable cameras), visual concepts (automatically extracted visual concepts with varying rates of accuracy), semantic content (semantic locations, semantic activities) based on sensor readings (via the Moves App) on mobile devices, biometrics information (heart rate, galvanic skin response, calorie burn, steps, continual blood glucose, etc.), music listening history, computer usage (frequency of typed words via the keyboard and information consumed on the computer via ASR of on-screen activity on a per-minute basis).

The training set has ten queries whose titles are described as: (1) Ice cream by the Sea, (2) Having food in a restaurant, (3) Watching videos, (4) Photograph of a Bridge, (5) Grocery shopping, (6) Playing a Guitar, (7) Cooking, (8) Car Sales Showroom (9) Public transportation, and (10) Paper or book reviewing.

The test set also has ten queries whose titles are listed as: (1) In a toyshop, (2) Driving home, (3) Seeking food in a fridge, (4) Watching football, (5) Coffee time, (6) Having breakfast at home, (7) Having coffee with two people, (8) Using smartphone outside, (9) Wearing a red plait shirt, and (10) Having a meeting in China.

It should be noted that each query has its additional descriptions where more strict rules are explained to confirm the content, concept, and context of the query.

The evaluation metrics are defined by imageCLEFlifelog 2019 as follows:

Algorithm 2 Cluster

Input: $I = \{I_i\}_{i=1..N}, F = \{(Vc_i, Va_i)\}_{i=1..N}, \theta_a, \theta_b$
Output: $\{C_k\}$

$k \longleftarrow 0$
$C = \{C_k\} \longleftarrow \emptyset$
$I^{temp} \Leftarrow SORT_{bytime}(I, F)$
repeat
  $v_a \longleftarrow I^{temp}.F.Va[0]$
  $v_c \longleftarrow I^{temp}.F.Vc[0]$
  $C[0] \Leftarrow \cup I^{temp}.I[0]$
  $I^{temp} \Leftarrow I^{temp} - I^{temp}[0]$
  for i=1..$\left\|I^{temp}\right\|$ **do**
    if $(similarity(v_a, I^{temp}.F.Va[i]) > \theta_a$ **and** $similarity(v_c, I^{temp}.F.Vc[i]) > \theta_c$ **then**
      $C[k] \Leftarrow \cup I^{temp}.I[i]$
      $I^{temp} \Leftarrow I^{temp} - I^{temp}[i]$
    else
      $k \longleftarrow k + 1$
      Break
    end if
  end for
until $\left\|I^{temp}\right\| == 0$
return C

- Cluster Recall at X (CR@X): a metric that assesses how many different clusters from the ground truth are represented among the top X results,
- Precision at X (P@X): measures the number of relevant photos among the top X results,
- F1-measure at X (F1@X): the harmonic mean of the previous two.

Figure 2 illustrates one example when applying the proposed method for the testing query 3: Seeking food in a fridge.

TABLE I
EVALUATION OF ALL RUNS ON THE TRAINING SET

Event	Run 1			Run 2		
	P@10	CR@10	F1@10	P@10	CR@10	F1@10
1	1.00	0.75	**0.86**	1.00	0.50	0.67
2	0.80	0.44	**0.57**	0.50	0.15	0.23
3	0.80	0.19	**0.31**	0.80	0.19	**0.31**
4	1.00	0.02	0.03	1.00	0.02	**0.04**
5	0.30	0.56	**0.39**	0.40	0.22	0.28
6	0.33	0.50	0.40	0.33	1.00	**0.50**
7	1.00	0.18	**0.30**	1.00	0.15	0.26
8	0.50	0.86	0.63	0.50	1.00	**0.67**
9	0.60	0.07	**0.13**	0.60	0.07	0.12
10	0.29	0.42	**0.34**	0.43	0.17	0.24
Avg	**0.66**	**0.40**	**0.40**	**0.66**	**0.35**	**0.33**

B. Evaluation and Comparison

Although three runs were submitted to evaluate, two best runs are chosen to introduce and discuss in this paper: (1) run 1: *interactive mode*. In this run, interactive functions are activated to let users interfere and manually get rid of those images that do not relevant to a query, (2) run 2: *automatic mode*. In this run, a program runs without any interfere from users.

TABLE II
EVALUATION OF ALL RUNS ON THE TEST SET

Event	Run 1			Run 2		
	P@10	CR@10	F1@10	P@10	CR@10	F1@10
01	1.00	0.50	**0.67**	1.00	0.50	**0.67**
02	0.00	0.00	0.00	0.00	0.00	0.00
	0.70	0.05	**0.09**	0.70	0.05	**0.09**
03	1.00	0.22	**0.36**	0.60	0.17	0.26
04	1.00	0.25	**0.40**	0.50	0.25	0.33
05	1.00	0.11	0.20	1.00	0.22	**0.36**
06	0.60	0.11	**0.19**	0.30	0.11	0.16
07	0.90	1.00	**0.95**	0.90	1.00	**0.95**
08	0.70	0.33	**0.45**	0.30	0.33	0.32
09	0.00	0.00	0.00	0.00	0.00	0.00
	0.90	0.14	**0.25**	0.50	0.14	0.22
10	0.70	0.33	**0.45**	0.70	0.33	**0.45**
Avg	0.69	0.29	0.37	0.53	0.29	0.35
	0.85	**0.31**	**0.40**	**0.65**	**0.31**	**0.38**

TABLE III
COMPARISON TO OTHERS

Event ID	F1@10								
	HCMUS	ZJUTCVR	BIDAL	DCU	ATS	REGIMLAB	UPB	TUC_MI	UAPT
1	**1.00**	0.95	0.67	0.57	0.46	0.44	0.29	0.46	0.00
2	**0.42**	0.17	0.09	0.09	0.09	0.06	0.08	0.13	0.00
3	0.36	0.40	0.36	0.29	0.36	**0.43**	0.07	0.00	0.07
4	**0.86**	0.37	0.40	0.67	0.00	0.40	0.00	0.00	0.00
5	**0.68**	0.47	0.20	0.36	0.36	0.54	0.34	0.00	0.11
6	0.00	**0.35**	0.19	0.17	0.26	0.00	0.00	0.18	0.14
7	0.89	**1.00**	0.95	0.75	0.33	0.00	0.18	0.00	0.00
8	0.32	0.00	**0.45**	0.00	0.00	0.00	0.00	0.00	0.00
9	**0.58**	0.44	0.25	0.00	0.12	0.00	0.00	0.00	0.00
10	**1.00**	0.25	0.45	0.00	0.57	0.00	0.32	0.40	0.25
Avg	**0.61**	**0.44**	**0.40**	**0.29**	**0.25**	**0.19**	**0.13**	**0.12**	**0.06**

Table I and II show the results running on the training and testing sets, respectively. That could lead to the conclusion that the proposed method probably is stable and robust enough to cope with different types of queries.

In all cases, the P@10 results are very high. It proves that the approach used for querying seeds of the proposed method is useful and precise. Moreover, the watershed-based stage after finding seeds can help not only to decrease the complexity of querying related images but also to increase the accuracy of event boundaries. Unfortunately, the CR@10 results are less accuracy comparing to P@10. The reason could come from merging clusters. Currently, clusters gained after running watershed are not merged and rearranged. That could lead to low accuracy when evaluating CR@X. This issue is investigated thoroughly in the future.

Nevertheless, misunderstanding context and content of queries sometimes lead to the worst results. For example, both runs failed in query 2 "driving home" and query 9 "wearing a red plaid shirt." The former was understood as "driving from office to home regardless of how many times stop at in-middle places," and the latter was distracted by synonym words of "plaid shirt" when leveraging GoogeSearchAPI to augmented the BoW. The first case should be understood that "driving home from the last stop before home," and the second case should focus on only "plaid shirt" not "sweater" nor "fannel." After fixing these mistakes, both runs have higher scores on query 2 and query 9, as described in Table II. The second rows of event 2, 9, and the average score show the results after correcting the mentioned misunderstanding. Hence, building a flexible mechanism to automatically build a useful taxonomy from a given query to avoid these mistakes is built in the future.

Nine teams participated to Task 2 included (1) HCMUS, (2) ZJUTCVR, (3) BIDAL (ourselves), (4) DCU, (5) ATS, (6) REGIMLAB, (7) TUCMI, (8) UAPT, and (9) UPB. The detail information of these teams could be referred to in [16]. Table III denotes the comparison among these teams. We are ranked in the third position. In general, the proposed method can find all events with acceptance accuracy (i.e., no event with zero F1@10 scores comparing to others those have at least one event with zero F1@10 scores). That confirms again the stability and anti-bias of the proposed method.

IV. CONCLUSIONS

We introduce the interactive watershed-based approach for lifelog moments retrieval that based on the hypothesis that if we can find a seed (i.e., an image that highly satisfied a given query), we can build a watershed by gathering its spatiotemporal neighbors that share the same content and context with the seed. Merging, splitting, and rearranging these watersheds generate a need-to-be-found lifelog moments satisfied the given query. The proposed method is thoroughly evaluated by the benchmark dataset provided by the image-CLEFlifelog 2019, and compared with other solutions coming from different teams. We also discuss the obstacle of misinterpreting a query leading to having a wrong answer. Hence, the interactive schema can be a good supporter for verifying the correctness of some stages of the proposed method. In the future, two following issues should be investigated to improve the productivity of the proposed method: (1) seed finding functions and watershed boundaries, and (2) a taxonomy generated automatically from queries.

Fig. 2. An Example of Using the Proposed Method for Querying

REFERENCES

[1] M. S. Dao, D. Nguyen, D. Tien, M. Riegler, and C. Gurrin, "Smart lifelogging: recognizing human activities using phasor," 2017.

[2] M. Dimiccoli, A. Cartas, and P. Radeva, "Activity recognition from visual lifelogs: State of the art and future challenges," in *Multimodal Behavior Analysis in the Wild*, pp. 121–134, Elsevier, 2019.

[3] M. Harvey, M. Langheinrich, and G. Ward, "Remembering through lifelogging," *Pervasive Mob. Comput.*, vol. 27, pp. 14–26, Apr. 2016.

[4] E. Kavallieratou, C. R. del Blanco, C. Cuevas, and N. García, "Retrieving events in life logging,"

[5] F. B. Abdallah, G. Feki, M. Ezzarka, A. B. Ammar, and C. B. Amar, "Regim lab team at imageclef lifelog moment retrieval task 2018,"

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[8] J. Lin, A. G. del Molino, Q. Xu, F. Fang, V. Subbaraju, J.-H. Lim, L. Li, and V. Chandrasekhar, "Vci2r at the ntcir-13 lifelog-2 lifelog semantic access task," 2017.

[9] L. Zhou, L. Piras, M. Riegler, G. Boato, D. Nguyen, D. Tien, and C. Gurrin, "Organizer team at imagecleflifelog 2017: baseline approaches for lifelog retrieval and summarization," 2017.

[10] T.-H. Tang12, M.-H. Fu, H.-H. Huang, K.-T. Chen, and H.-H. Chen13, "Visual concept selection with textual knowledge for understanding activities of daily living and life moment retrieval,"

[11] M.-T. Tran, T. Dinh-Duy, T.-D. Truong, V.-K. Vo-Ho, Q.-A. Luong, and V.-T. Nguyen, "Lifelog moment retrieval with visual concept fusion and text-based query expansion,"

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[13] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[14] G. Patterson, C. Xu, H. Su, and J. Hays, "The sun attribute database: Beyond categories for deeper scene understanding," *Int. J. Comput. Vision*, vol. 108, pp. 59–81, May 2014.

[15] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.

[16] D.-T. Dang-Nguyen, L. Piras, M. Riegler, M.-T. Tran, L. Zhou, M. Lux, T.-K. Le, V.-T. Ninh, and C. Gurrin, "Overview of ImageCLEFlifelog 2019: Solve my life puzzle and Lifelog Moment Retrieval," in *CLEF2019 Working Notes*, CEUR Workshop Proceedings, (Lugano, Switzerland), CEUR-WS.org <http://ceur-ws.org>, September 09-12 2019.