Smart Lifelog Retrieval System with Habit-based Concepts and Moment Visualization

Tokinori Suzuki[1] and Daisuke Ikeda[1]

[1] Department of Informatics, Kyushu University, Fukuoka, Japan
suzuki.tokinori.070@s.kyushu-u.ac.jp, daisuke@inf.kyushu-u.ac.jp

Abstract. Our QUIK team participated in the Lifelog Semantic Access Subtask (LSAT) of the NTCIR-14 Lifelog-3 task. The task is that, given a topic of users' daily activity or events (e.g. Find the moment when a user was taking a train from the city to home) as a query, a system retrieves the relevant images of the moments from users' images of recording their daily lives. For LSAT task, we present an approach to retrieve users' lifelog images by computing the similarity between users' lifelog images and images obtained from the web by querying the LSAT topics into a web search engine. For computing the similarity between the lifelog images and images from the web, we employ a classifier trained on the images collected from the web with a convolutional neural network model. This paper describes our approach to solving LSAT task and reports the official results that we got.

Team Name. QUIK

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

Keywords: Lifelog, Multimedia, Image Recognition.

1 Introduction

The QUIK team participated in the Lifelog Semantic Access Subtask (LSAT) of the NTCIR-14 Lifelog-3 task [3]. In LSAT task, participants are required to retrieve a number of specific moments, which are defined as events or activities happened throughout a day, in a lifelogger's life. Specifically, given a query topic which represents an event or activity in users' daily life (e.g. Find the moment when a user was taking a train from the city to home), a system retrieves relevant moments from users' lifelog images and the associated meta information such as location, timestamp, the number of their steps and so on.

This Lifelog-3 task is the third time of the series of NTCIR Lifelog task since NTCIR-12 [1]. Participants in the previous NTCIR13 Lifelog-2 LSAT task [2] mentioned that the central difficulty of the task is on understanding users' events or activities from lifelog images with the meta data. In relation to image-related task, while Object Recognition (OR) has been actively studied in this decade such as in ILSVRC

competition [8], the difficulty of LSAT is different from OR. OR is a task to classify objects depicted in an image into pre-defined classes. For example, OR may classify an image of a user's commuting on a train into "train interior" or "person" classes. Compared to this type of task, LSAT is further complex task because if we have objective information on images, we still need to link the objective and other visual information to users' event or activity (i.e., query topics). Taking the running example, a system is required to link the visual information such as OR labels to the query of commuting on a train. Usually, there are no direct connection between lifelog images and query topics representing events/activities (i.e., semantic gap) [5,10].

To solve the LSAT task, we made an attempt to retrieve the relevant moments only on visual features of the lifelog data, i.e., lifelog images. Our proposed approach is based on the assumption that the relevant lifelog images for a query topic are similar to images on those visuals obtained by querying the topic into a web search engine. For example, we expect that there can be a similarity between users' lifelog images recorded when they were on a train and many pictures which can be obtained by querying a LSAT query topic (e.g., "I am taking a train from the city to home") into a web search engine. So, we collect the images as external data from the web and train a classifier with a convolutional neural network model.

The reminder of this paper is organized as follows. Section 2 explains the proposed approach. Section 3 presents the official results of our approach and discussions. Section 4 concludes the paper briefly.

2 Proposed Approach

We proposed a retrieval approach using query topic classification for the LSAT task. Fig. **1** displays the framework of the approach. The approach mainly consists of two components of similarity computing on 1) visual concepts and 2) the query topic similarity between lifelog images and a query. In the rest of this section, we explain each components of the approach.

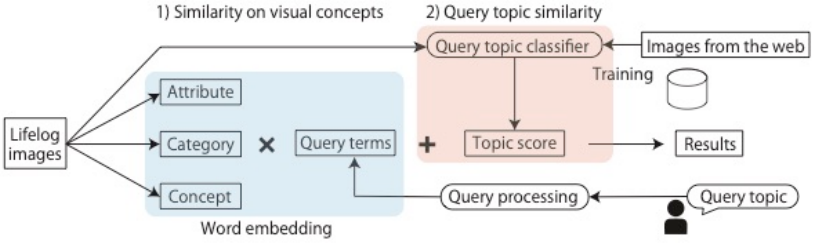

Fig. 1. Framework of our proposed approach

2.1 Visual Concepts of Images

In the LSAT task, participants are given visual concept information associated with users' lifelog images. We use the visual concept information as the basis of our approach on 1) similarity on visual concepts. The Lifelog-3 task organizers prepared three types of concept information called attributes, categories and concepts. These information are automatically generated for each image.

Fig. 2 shows an example of visual concepts of an image. First, attributes represent for the information of environmental scenes, places or objects in images. In the figure, the attribute includes, for example, enclosed area, indoor lighting, studying and so on as the labels. Ten attribute labels are assigned to each image. Second, categories are for places relevant to an image. The place information, e.g. home office, office, computer room and so on, are given as top five similar place labels for an image along with the scores by the classifier. Third, concepts represent for objects in images. For example, the labels are chair, laptop and keyboard in the figure. The number of the concept label is as much as the classifier can recognize the objective concepts in an image up to 25 labels each an image. When the classifier cannot recognize the trained objects in the images, a NULL label is assigned to the image. The concept information includes the bounding box of the object location within the image and the classification scores.

Attribute	Category	Cat. score	Concept	Con. score
enclosed area	home office	0.595	chair	0.843
indoor lighting	office	0.206	laptop	0.987
studying	comput. room	0.076	keyboard	0.866
...	...	...	NULL	–

Fig. 2. Example of visual concepts of a lifelog image

2.2 Collecting Query Topic Pictures and Training Classifier

We, then, explain how to obtain 2) query topic similarity in Fig. 1 by a topic classifier. In order to train the topic classifier, we collected pictures representing query topics from the web. Querying modified title or description part of each topic into Google image search[1], we obtained around 400 images per topic. The queried title of a topic, for example, is "I am eating icecream beside the sea" from the original topic title "Find the moment when a user was eating icecream beside the sea". For some of the topics, e.g. topic id 11, 15 and 21, with only one word for those title, we used the description to generate search words with the same manner. Since the images are obtained from the web and noisy for representing the topics, we conducted human judgement whether the pictures relevant to the topics or not. Finally, we prepared averagely 170 images for each topic. The detail numbers of the collected images are shown in Table 1. We note that since the web pictures which seem relevant to topics are small on the some of the topics, we could prepare only around 20 pictures on topics id 14014, 14020 and 14022.

[1] https://images.google.co.jp/

4

For training the topic classifier, we use the *convolutional neural network* (CNN) model of *deep residual network* with 50 layers (Resnet-50) [4]. Splitting the collected images into 70% for training and 30% for validation each a query topic, we train the classifier with Resnet-50 model. We set the hyper parameters of the model with batch size at 32 and learning rate at 0.0001. The validation accuracy of the trained classifier was 0.80. We use the probabilities of the softmax function as the scores.

Table 1. Collectd images for each topic

Topic	# of images	Topic	# of images	Topic	# of images
14001	117	14009	183	14017	136
14002	259	14010	357	14018	109
14003	232	14011	383	14019	40
14004	191	14012	211	14020	22
14005	245	14013	220	14021	164
14006	256	14014	29	14022	21
14007	216	14015	313	14023	47
14008	179	14016	69	14024	62

2.3 Similarity Computing

The visual concepts are given in the form of words. If keyword-based retrieval is applied to visual concept labels, we aware that there is a gap between vocabularies of visual concept labels and words in queries, and it could lead the mismatch of keyword searches. In addition, query topics representing lifelog activities and events may not have direct links to visual concepts of images. For example, when a query topic is "Find the moments when a user was eating any food at his/her desk at work", the visual concepts of the expected images may be labeled with "home office" or "office" for the category labels, and with a "banana" label for the concept label. Thus, the same words or phrases in query topics do not appear in the visual concepts of images. To overcome the difference of the vocabularies and measure the similarity between query topic and visual concepts, we compute the similarities over distributed representation of words learned by word embedding [7].

To obtain vector representation of words for visual concepts and queries, we trained the word embedding with skip-gram model on English version of Wikipedia dump data on October 2017. We use the learned 200 dimensional vectors for the words.

We define the similarity of two words w_1 and w_2 as the cosine similarity between the learned word embeddings:

$$sim(w_1, w_2) = \cos(\boldsymbol{w}_1, \boldsymbol{w}_2) = \frac{\boldsymbol{w}_1^T \boldsymbol{w}_2}{\|\boldsymbol{w}_1\| \|\boldsymbol{w}_2\|}, \quad (1)$$

where the cosine similarity is calculated by the inner product of vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. Next, to compute the similarity between two *bags-of-words,* query topic Q and visual

concepts V, we use text-to-text similarity introduced by Mihalcea et al. [6]. The similarity between a word w_q in Q and the set of words in V is computed as the maximum similarity between w_q and any word w_v in V:

$$sim(w_q, \mathrm{V}) = \max_{w_v \in V} sim(w_q, w_v) \quad (2)$$

Then, we compute the global similarity between query topic Q and visual concepts V of image I by the sum of cosine similarity and the topic score as follows:

$$sim(\mathrm{Q}, \mathrm{V}) = \sum_{q \in Q} sim(q, V) + topic\ score(I) \times \alpha, \quad (3)$$

where $topic\ score(I)$ is the probability of an image I for the query topic classified by the classifier explained in Section 2.1, α is a weighting parameter for the score. We set α at 1.5 throughout the experiment.

3 Experiment

In this section, first we mention to NTCIR-14 Lifelog-3 test collection that we evaluated our approach on. Then, we introduce the official results of our approach.

3.1 Test collection and Query Processing

NTCIR-14 Lifelog-3 test collection consists of daily pictures recorded by two users (user 1 and user 2) and associated retrieval topics.
istinguish the types of images.

Table 2 summarizes the image statistics of the test collection. In the test collection, there are 64,132 and 17,615 lifelog images for user 1 and user 2, respectively, during the period from 3 May 2018 to 31 May 2018. The test collection images include two types of pictures. First type of pictures is passively captured by the wearable camera (OMG autographer[2]) clipped to their clothing or lanyard around neck from the users' view point. Second type of pictures is manually shot by the users. In this experiment, we did not distinguish the types of images.

The collection has 24 topics in total, 16 topics targeted for user 1, five topics targeted for user 2 and three topics targeted for both of the users. We make query terms for each topic using *part-of-speech* (POS) information of the query topics. We analyzed POS of the title text of query topics with Stanford POS Tagger [9], and set words only with a noun or a verb label as the query terms for the topic. As a result, we prepared average 2.6 query terms for each topic.

[2] https://en.wikipedia.org/wiki/Autographer

6

Table 2. NTCIR-14 Lifelog-3 test collection statistics

User	Period	The # of days	The # of images
User 1	3 May 2018~ 31 May 2018	29 days	64,132
User 2	9 May 2018 ~ 22 May 2018	14 days	17,615
Total		43 days	81,747

3.2 Approach Feature Used in Two Runs

We submitted two runs of our approach varied on the features used in similarity computation. Table 3 summarizes the features of the two runs. The difference among the runs is on using query topic similarity so that we evaluate the effectiveness of the feature. Run 1 used visual concepts only, and Run 2 used both of the visual concepts and the query topic similarity.

Table 3. Features used in each run

Run	Attribute	Category	Concept	Query topic
Run 1	✔	✔	✔	
Run 2	✔	✔	✔	✔

3.3 Results

We submitted two runs of our approach, and received the evaluated results. The official results are shown in Table 4.

The table summarizes the mean average precision (MAP), precision at fifth, tenth and 30th (P@5, P@10, P@30, respectively) of the two runs over all the topics. The MAP of both runs are 0.0559 at Run1 and 0.0544 at Run 2, which is not very different from each other. The precision measures show the close values between two runs too. Fig. 3 shows the average precision (AP) of each topic. Except for the two topics, 14001 and 14023, the values are lower than 0.12 including 0 of AP on ten topics. As a result, the AP on these topics decreased the MAP.

We mention to the effectiveness of the query topic similarity. Even though the MAP of Run 2 using the query topic similarity is lower than Run 1, which does not use the query topic similarity, AP of the runs fluctuate on topics as seen in Fig. 3., thus the additional feature can improve the retrieval on some of the topics. The first reason about the quite low improvement of Run 2 would be that we tested the approach with only one parameter setting of the α parameter in equation (3). There might be the better setting. The second reason is about the training data about the topic classifier. We could not collect small number of the training images on several topics.

Our approach could not retrieve any relevant results on ten topics in Fig. 3. We think that this may be due to using query terms only in the title part of the topics, which is more abstract of information needs than the query description and narrative.

Table 4. Official results of two runs. Each measure indicates the average over all the topics.

Run	MAP	P@5	P@10	P@30
Run 1	0.0559	0.1583	0.1583	0.1181
Run 2	0.0544	0.1583	0.1458	0.1194

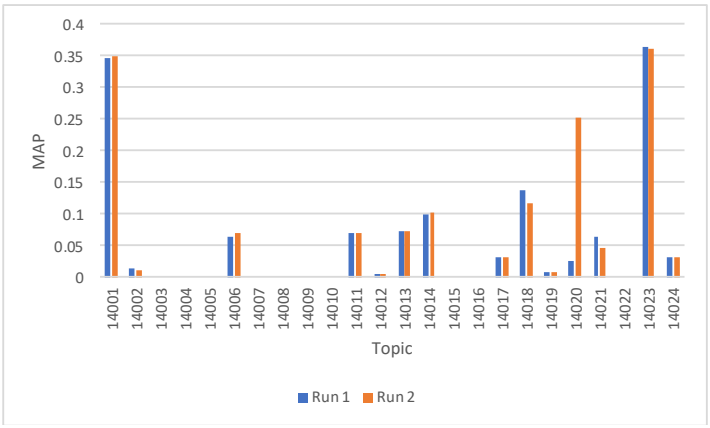

Fig. 3. AP of two runs each topic

4 Conclusion

In this paper, we present an approach for NTCIR-14 Lifelog-3 Lifelog Semantic Access Subtask (LSAT). Our approach retrieves the relevant lifelog images by computing similarity between lifelog images and images obtained by querying a topic into a web search engine. We report the official results.

On the some of the topics, the approach shows a certain level of the retrieval performance. Though, the approach cannot retrieve the relevant results on about the half of the topics, possibly due to query terms used in the experiments. An immediate future work is trying to use words in the description or narrative text of the topics for the retrieval.

References

1. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Albatal, R.: Overview of ntcir-12 lifelog task. In: *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*. pp. 354–360 (2016).
2. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Gupta, R., Albatal, R. and Dang-Nguyen, D.: Overview of NTCIR-13 Lifelog-2 Task. In: *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*. pp. 6–11 (2017).
3. Gurrin, C., Joho, H., Hopfgartner, F., Dang-Nguyen, D., Zhou, L., Ninh, V., Le, T., Albatal, R. and Healy, G.: Overview of NTCIR-14 Lifelog-3 Task. In: *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*. (2019)

8

4. He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition. In: *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778 (2016).
5. Lin, J., Garcia del Molino, A., Xu, Q., Fang, F., Subbaraju, V. and Lim, J.: VCI2R at NTCIR-13 Lifelog-2 Lifelog Semantic Access Task. In: *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*. pp. 28–32 (2017).
6. Mihalcea, R., Corley, C. and Strapparava, C.: Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In: *Proceedings of the 21st National Conference of Artifical Intelligence*. pp. 775–780 (2006).
7. Mikolov, T., Sutskever I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems*. pp. 3111–3119 (2013).
8. Russakovsky, O., Deng, J., Su H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115, 3, pp. 211–252 (2015).
9. Toutanova, K., Klein, D., Manning, C. and Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. pp. 252–259 (2003).
10. Yamamoto, S., Nishimura, T., Akagi, Y., Takimoto, Y., Inoue, T. and Toda, H.: PBG at NTCIR-13 Lifelog-2 LAT, LSAT, and LEST Tasks. In: *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*. pp. 12–19 (2017).