An Interactive Approach to Integrating External Textual Knowledge for Multimodal Lifelog Retrieval

Chia-Chun Chang[1], Min-Huan Fu[1], Hen-Hsen Huang[2,3] and Hsin-His Chen[1,3]
[1] Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
[2] Department of Computer Science, National Chengchi University, Taipei, Taiwan
[3] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan
{ccchang, mhfu}@nlg.csie.ntu.edu.tw;
hhhuang@nccu.edu.tw; hhchen@ntu.edu.tw

ABSTRACT

The semantic gap between textual queries and visual concepts is one of the key challenges in lifelog retrieval. This work presents an interactive system aimed at improving the retrieval accuracy by query term suggestion. Besides, this system also assists users to refine the retrieval results by image similarity clustering. For recommending a list of candidate words, we extract visual concepts from images by using computer vision models, and then incorporate both official and additional concepts into our system using pre-trained word embedding, in which textual knowledge is inherent. We also purpose an intelligent mechanism for rapidly removing multiple irrelevant search results. For reaching out this purpose, we build kd-trees [1] offline for reducing the computational overhead and cluster similar images by nearest neighbor search in the embedding space. Whenever users exclude some irrelevant images, their nearest neighbors in the image embedding space are also removed. In this way, users can efficiently screen out the relevant results and purge the irrelevant ones, scanning over more retrieval results in a shorter period of time.

KEYWORDS

Lifelog, Interactive System, Word Embedding, Multimodal Learning

ACM Reference Format:

Chia-Chun Chang, Min-Huan Fu, Hen-Hsen Huang and Hsin-Hsi Chen. 2019. An Interactive Approach to Integrating External Textual Knowledge for Multimodal Lifelog Retrieval. In *LSC'19: The Lifelog Search Challenge, June 10, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3326460.3329163

1 INTRODUCTION

Lifelogging analysis becomes extremely important as lifelog data has been tremendously growing and the high availability of wearable lifelogging devices has been offering a novel choice for daily life recording. Using lifelogging devices to capture daily life provides a rich resource for daily life understanding and memory recall. Generally, huge volume of multimedia archives is expected to be generated for analytical and retrieval purpose as time proceeds. Efficient approaches for users to organize and access collected lifelog data are thus highly demanded. In the Lifelog Search Challenge 2019 (LSC 2019), the organizers provide a dataset [2] which contains data captured via wearable devices from a continuous time span of about 4 weeks. Moreover, the organizers provide an evaluation system to assist participants to develop their lifelog interactive systems before the workshop. In this evaluation system, only a few clues are issued in the beginning, and more clues are shown to participants incrementally as time goes on. The LSC server that issues query cues and collects the responses of systems will announce the correct result after 5 minutes for a query. Participators should input image IDs after obtaining retrieval results as soon as possible.

In this task, we seek to reduce the semantic gap between textual queries and visual concepts by bringing into the semantic word embedding into our retrieval scheme. Moreover, based on the experiences of our previous works [3] in ImageCLEF 2018, there are two main issues to be tackled while formulating an appropriate query for a given topic. Firstly, it is difficult for users who are not familiar with the lifelog dataset, a personal big data, to choose adequate query words related to each topic. To this end, our system automatically suggests users a list of query words for concept word selection. Secondly, the system tends to retrieve batches of visually similar images, since they all contain similar visual concepts. An efficient way for deleting irrelevant results thus plays an important role for diversifying the retrieval results.

To optimize the searching efficiency when users submit query, we propose an interactive system that facilities two functions for users to enhance the retrieval outcome: (1) Our system automatically suggests the query words by returning top k similar concept words. (2) For irrelevant results, our system provides a mechanism that excludes similar images by calculating the nearest neighbors exhaustively in the embedding space.

LSC'19, June 10, 2019, Ottawa, ON, Canada
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6781-3/19/06...$15.00
https://doi.org/10.1145/3326460.3329163

2 DATA PREPROCESSING

2.1 An Overview

The lifelog dataset provided by the LSC organizers consists of 27 days of data from an individual lifelogger, collected with the Narrative Clip 2. The multimodal dataset includes images, biometrics, and human activities. Figure 1 shows an overview of the offline preprocessing. It is composed of two major steps to preprocess lifelog data – say, image preprocessing and visual concept labeling.

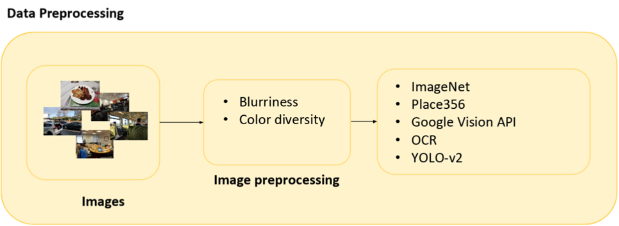


Fig. 1. An overview of offline data preprocessing.

2.2 Image Preprocessing

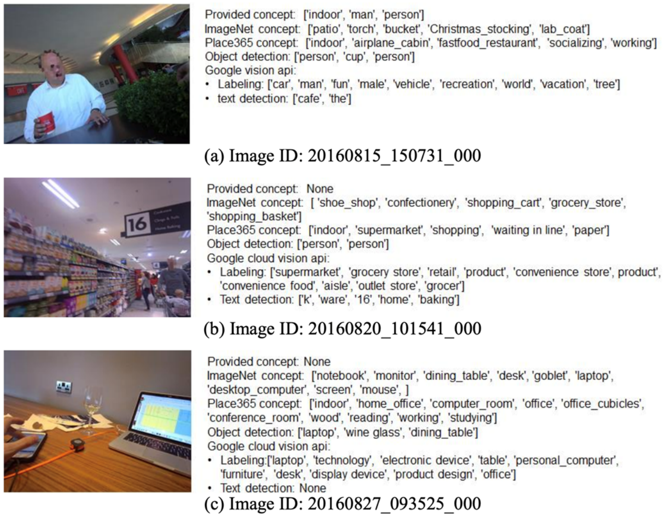
Images in the lifelog dataset are automatically captured via wearable recording device. Lifelogger may perform different activities in his/her daily life, so that images may suffer from the poor quality due to overexposed, unexpected shaking or out of focus. The dataset contains a number of blurry or occluded images captured with wearable cameras. We observe that the advanced CV models are prone to error for images with poor quality. To ensure the quality of input images to our models, we apply blurriness and color diversity detection [4] during preprocessing stage.

We prune low quality images with blurriness and color diversity detection. The blurriness metric is defined based on the variation of the Laplacian. Each image is convolved with a 3×3 Laplacian filter, and the blurriness measurement is calculated as the variance of the convolved result. Images with low variance are considered blurry and undesirable. Besides, images with high color homogeneity are also considered uninformative, and can be detected with quantized color histograms.

2.3 Visual Concept Labeling

Our previous work [3] has shown that effective textual representation for images is crucial to lifelog retrieval. Given an image, we would like to retrieve as more useful information as possible. As key components for describing life events, we would like to know where a photo was taken, what objects there were, and even what actions the lifelogger took at that moment. We adopt DenseNet [5], one of the current state-of-the-art deep neural models, to extract visual concepts from different aspects. In general, we detect visual semantic concepts and scene of images using two DenseNet classifiers, which are pre-trained on ImageNet1K [6] and Place365 [7], respectively. The classes with output probability beyond a given threshold are chosen as labels. It is kept moderate to ensure the recall rate. We refer to the setting in [3].

In order to detect the detail presenting in images, we apply object detection models to each image, including Yolo-v2 [5] and Faster RCNN [8]. Both models are pre-trained on MS-COCO [9] and Open Images [10] datasets. MS-COCO dataset contains 91 objects types with 2.5M labels in 328K images. Open Images dataset, on the other hand, consists of about 15M boxes on 600 categories. Both datasets cover most of the domains close to human daily life. We also exploit the image analysis tools provided by Google Cloud Vision API[1]. The online service provides not only fruitful visual labels but also supports optical character recognition (OCR), which helps detect and extract text from images. Finally, we include the image concepts provided by the LSC organizers. After going through the above process, an image would be tagged with concepts present in various aspects, as shown in Fig. 2.



Provided concept: ['indoor', 'man', 'person']
ImageNet concept: ['patio', 'torch', 'bucket', 'Christmas_stocking', 'lab_coat']
Place365 concept: ['indoor', 'airplane_cabin', 'fastfood_restaurant', 'socializing', 'working']
Object detection: ['person', 'cup', 'person']
Google vision api:
- Labeling: ['car', 'man', 'fun', 'male', 'vehicle', 'recreation', 'world', 'vacation', 'tree']
- text detection: ['cafe', 'the']

(a) Image ID: 20160815_150731_000

Provided concept: None
ImageNet concept: ['shoe_shop', 'confectionery', 'shopping_cart', 'grocery_store', 'shopping_basket']
Place365 concept: ['indoor', 'supermarket', 'shopping', 'waiting in line', 'paper']
Object detection: ['person', 'person']
Google cloud vision api:
- Labeling: ['supermarket', 'grocery store', 'retail', 'product', 'convenience store', product', 'convenience food', 'aisle', 'outlet store', 'grocer']
- Text detection: ['k', 'ware', '16', 'home', 'baking']

(b) Image ID: 20160820_101541_000

Provided concept: None
ImageNet concept: ['notebook', 'monitor', 'dining_table', 'desk', 'goblet', 'laptop', 'desktop_computer', 'screen', 'mouse',]
Place365 concept: ['indoor', 'home_office', 'computer_room', 'office', 'office_cubicles', 'conference_room', 'wood', 'reading', 'working', 'studying']
Object detection: ['laptop', 'wine glass', 'dining_table']
Google cloud vision api:
- Labeling:['laptop', 'technology', 'electronic device', 'table', 'personal_computer', 'furniture', 'desk', 'display device', 'product design', 'office']
- Text detection: None

(c) Image ID: 20160827_093525_000

Fig. 2. Lifelog images with their corresponding visual concepts labeled by various image recognition models.

3 INTERACTIVE RETRIEVAL SYSTEM

Generally, there are a plenty of diversified ways to retrieve lifelog data, such as retrieving data based on modeling the similarity between textual concepts and user queries, or based on both visual and metadata information along with clustering results. Nevertheless, an interactive retrieval system should not only provide accurate retrieval results, but also offer an intuitive and simple operation. A better retrieval mechanism can retrieve more relevant results, but a good interactive system can make an impact on users’ query and result refinement. In the following section, we will introduce our retrieval mechanism and interactive operations.

[1] https://cloud.google.com/vision/

3.1 Retrieval Mechanism

In our retrieval mechanism, shown in Fig. 3, we represent lifelog images as short documents consisting of concept words, and are associated with the metadata recorded by lifelogging device, such as time information. Given a set of query words, we apply BM25 [11] as the ranking function for document retrieval. BM25 measures the probabilistic relevance between two sets of concept words based on term frequency (TF) and inverse document frequency (IDF). In this way, rare concepts are given more importance and are more likely to be captured in the retrieval scheme. The retrieval results are listed in the descending order of their ranking scores.

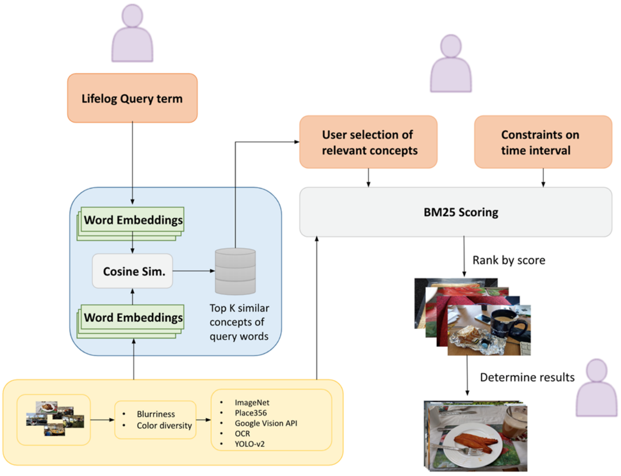


Fig. 3. Lifelog images retrieval interactive system

3.2 Interactive Operations

This section elaborates our interactive retrieval system in detail, showing how it benefits lifelog retrieval tasks. For allowing users to efficiently refine the retrieval results, the system is built as a search engine that provides a customized retrieval operation for users. Our previous work [3] considers that crucial components for lifelog understanding include the identification of what, where, and when a lifelogger does. The components can be covered by visual concepts of lifelog images along with the time information recorded by lifelogging devices. Accordingly, we provide two options in our system to allow users to specify the queries and the search constraints. We do not provide more options because the time of retrieval is limited.

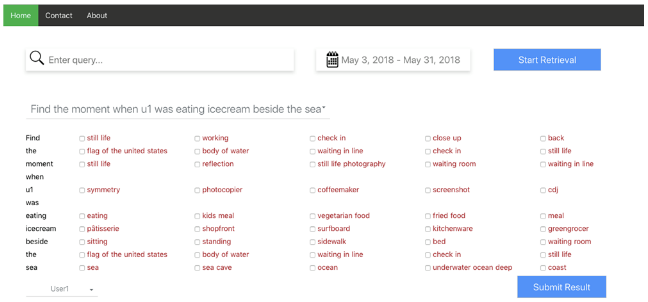
The whole process of our interactive system is as follows: typing query words, deciding time interval and selecting relevant images. The system automatically suggests users with a list of semantically related concept words to each query word, which will be illustrated in the next part. Users can choose the suggested query words, or manually input words as additional query terms. Finally, our system will return all images that match the given query words and constraints.

Query Suggestion. Our query expansion strategy is improved over our previous work [3], in which the retrieval system offers the suggested concept words related to each query. The main difference is that we restrict the search space of the nearest neighbor search to a very small range, including only the extracted concept words presented in the dataset. Due to the small vocabulary size of the extracted concept words, it is computationally feasible to search the nearest neighbors exhaustively in the embedding space. We exploit pre-trained word embeddings to obtain the top *k* (*k*=5 by default) similar concept words by comparing the semantic similarity between concept words and query terms.

For example, we conduct query expansion on each word of the sentence "find the moment when u1 was cooking food on a BBQ." And we filter out some prepositions, shown in Fig. 4. The suggestion results contain diversified words and phrases, such as barbecue, outdoor grill, and outdoor grill rack & topper.

Furthermore, we show the difference between queries with suggestion words and queries without suggestion words in NTCIR-14 dataset. In Fig. 5(a), we use the query word "BBQ" to retrieve images, and the retrieval result contains only 20 qualified images of the first 150 images. In Fig. 5(b), we conduct query expansion and get "barbeque" from BBQ's query expansion. Using "barbeque" as the query word, we obtain almost 142 qualified images of the first 150 results.

As another example shown in Fig. 6, we obtain "television" from TV's query expansion and try to use both of them as query words. Using "TV" as the query word only resulted in two qualified images out of the first 150 images, as shown in Fig. 6(a). However, 143 qualified images shown in Fig. 6(b) were obtained from the first 150 images when the expanded query "television" was used as the query word. In general, query expansion plays an important role for multimodal information retrieval by including more relevant results.



Find the moment when u1 was eating icecream beside the sea

Find	□ still life	□ working	□ check in	□ close up	□ back
the	□ flag of the united states	□ body of water	□ waiting in line	□ check in	□ still life
moment	□ still life	□ reflection	□ still life photography	□ waiting room	□ waiting in line
when					
u1	□ symmetry	□ photocopier	□ coffeemaker	□ screenshot	□ cdj
was					
eating	□ eating	□ kids meal	□ vegetarian food	□ fried food	□ meal
icecream	□ pâtisserie	□ shopfront	□ surfboard	□ kitchenware	□ greengrocer
beside	□ sitting	□ standing	□ sidewalk	□ bed	□ waiting room
the	□ flag of the united states	□ body of water	□ waiting in line	□ check in	□ still life
sea	□ sea	□ sea cave	□ ocean	□ underwater ocean deep	□ coast

Fig. 4. Query words recommended by query expansion.

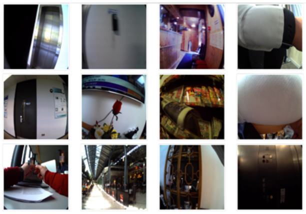


Fig. 5(a). Retrieve images with with query word "BBQ".

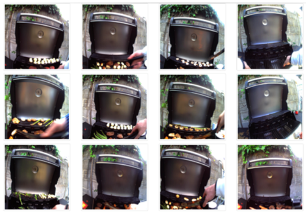
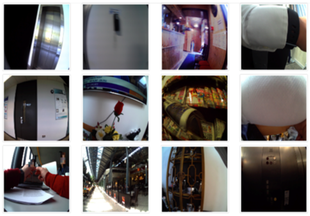


Fig.5(b). Retrieve images query word "barbecue".



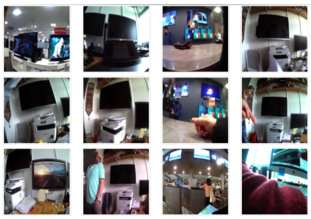


Fig. 6(a). Retrieve images with query word "TV". **Fig.6(b). Retrieve images with query word "television".**

Retrieval Result Refinement. As an interactive system, it is crucial to provide an efficient search mechanism to users for their information needs. By offering a simple refining process, our system allows users to remove irrelevant results or to pick relevant images by clicking right and left button on the images, respectively. Our system will assist users to reach out the purpose mentioned above by the following conditions: (1) Whenever users remove irrelevant images, our system will also automatically eliminate similar images by clustering results, shown in Fig. 7. (2) Users can pick up target images, or remove all irrelevant, and then click submit. The remaining images are considered relevant and submitted as the final result.

The process of automatic irrelevant result removal is described as follows. Initially, irrelevant image removal was designed as single deletion on each clicking. However, the manually process will be very inefficient for multiple irrelevant results. Since we apply BM25 as our ranking function, the retrieval results show in descending order of their ranking scores. The retrieval result may contain a number of uninformative but similar images, and manual deletion of these images is time-consuming. For this reason, we attempt to propose a mechanism that automatically eliminates other retrieval outcomes similar to the one manually deleted by the user. In the sense that we represent each image as a high-dimensional feature vector (embedding), clustering can be employed for nearest neighbor search in the embedding space. We build k-d trees offline for reducing the computational overhead for the nearest neighbor search. Whenever a user marks any retrieved image as irrelevant, our system automatically collects similar images to be removed by traversing the pre-built k-d tree.

4 CONCLUSION

In this paper, we propose an interactive retrieval system that allows users to retrieve their lifelog in a more efficient way. We also apply external textual knowledge to reduce the semantic gap between textual queries and the visual concepts extracted by computer vision models. With the recommended query words, users can infer more relevant queries to reduce time spent on thinking of relevant keywords. We also propose a deletion mechanism that searches the nearest neighbors in the image embedding space, providing users with an efficient way to delete images.

In future work, we plan to work on additional features that will rerank retrieval results and recommend the similar results when a user clicks on an image. In this way, the time spending on scanning over all images might be decreased, and users will experience a more convenient process.

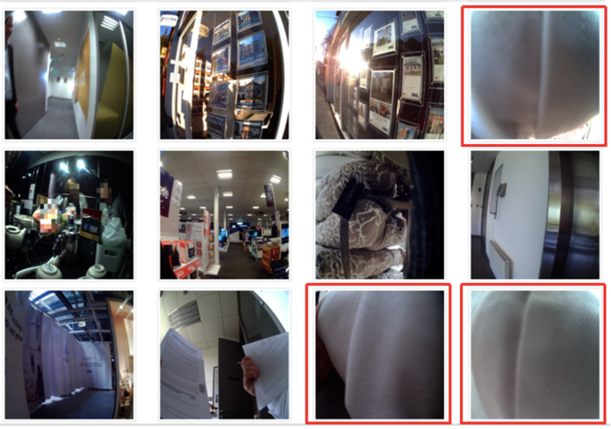


Fig. 7. Retrieval results containing high similarity irrelevant images.

ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-107-2634-F-002-011-, MOST-108-2634-F-002-008-, MOST-107-2218-E-009-050-, and AS-TP-107-M05.

REFERENCES

[1] Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. Communications of the ACM, 18(9), 509-517.

[2] Gurrin, Cathal , Schoeffmann, Klaus, Joho, Hideo, Munzer, Bernd, Albatal, Rami, Hopfgartner, Frank , Zhou, Liting and Dang-Nguyen, Duc-Tien (2018) A Test collection for interactive lifelog retrieval. In: MMM 2019, the 25th International Conference on MultiMedia Modeling, 8-12 January 2019, Thessaloniki, Greece.

[3] Tang, T.H., Fu, M.H., Huang, H.H., Chen, K.T., Chen, H.H.: Visual concept selection with textual knowledge for understanding activities of daily living and life moment retrieval. In: CLEF2018 Working Notes (CEUR Workshop Proceedings) (2018).

[4] Garcia del Molino, A., Mandal, B., Lin, J., Hwee Lim, J., Subbaraju, V., Chandrasekhar, V.:VC-I2R@ImageCLEF2017: Ensemble of Deep Learned Features for Lifelog Video Summarization. In proceedings of CLEF (2017)

[5] Huang, G., Liu, Z., Maaten, L., Weinberger, K.: Densely Connected Convolutional Networks. In proceedings of CVPR (2017).

[6] J., Deng, W., Dong, R., Socher, L.J., Li, K., Li, L., Fei-Fei: ImageNet: A LargeScale Hierarchical Image Database. In proceedings of CVPR (2009).

[7] Bolei, Z., Agata, L., Aditya, K., Aude, O., Antonio, T.: Places: A 10 Million Image Database for Scene Recognition. In proceedings of TPAMI (2017).

[8] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K.: Speed/accuracy trade-offs for modern convolutional object detectors. In proceedings of CVPR (2017).

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In proceedings of ECCV (2014).

[10] 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. (2017).

[11] Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. In: Foundations and Trends in Information Retrieval archive, Vol 3 Issue 4. ACM (2009).