Solving Life Puzzle with Visual Context-based Clustering and Habit Reference

Trung-Hieu Hoang[1], Mai-Khiem Tran[1], Vinh-Tiep Nguyen[2], Minh-Triet Tran[1]

[1] University of Science, VNU-HCM, Ho Chi Minh city, Vietnam
{hthieu, tmkhiem}@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. Lifelogging has taken a wide range of interest from research communities due to the increasing number of wearable and personal devices. With a large amount of data collected every day, an essential task is to organize, manage, and retrieve data efficiently. Therefore, from a collection of photos, it is necessary to rearrange them in some order that may represent a meaning sequence of events in daily life. This motivates the task of Life Puzzle in ImageCLEFlifelog 2019. To solve this task, we propose a novel method to exploit personal habits in their daily routine activities. First we group images into several clusters based on the their visual similarity and extracted visual concepts. For each image cluster, we utilize our Bag-of-Visual-Words framework to query similar scenes from personal lifelog data to predict the possible time instant in a day of the cluster. Using our proposed method, we achieve the first place at Multimedia retrieval in CLEF - “Solve my life puzzle” challenge with competitive Kendall's $\mathcal{T}$ score up to 0.4 and the final score of 0.55.

Keywords: Life puzzle· Meta-data image clustering · Image retrieval · Bag-of-Visual-Words · Root SIFT features

1 Introduction

Lifelog data is a valuable source of information to provide a better understanding of people's daily activities, habits, and behaviours. Everyday, each people may, intentionally or unintentionally, create a huge amount of useful lifelog data in various formats, such as photos, video clips, audio clips, text notes, or activities logs in computers or mobile devices. Among these data format, Visual data is one of the most important and interesting sources to analyze people's daily activities.

Sometimes, visual lifelog data are provided as a collection of photos without any specific temporal order. In this case, it is important to arrange the given photos in some meaning meaningful order that reflects a reasonable sequence of daily activities. This is the motivation for the “Solve my life puzzle” challenge in the ImageCLEFlifelog 2019 [5]. In this challenge, participants are given samples of lifelog data, which are comprised of images along with metadata (i.e biometrics and GPS location), and are tasked to re-organize the given data in an ascending timeframe order and predict four categories, with respect to four timespans of a day (morning, afternoon, night, mid-night).

In the ideal case, we aim to recover the original sequence of photos corresponding to the real sequence of events that actually happen in a day of a user. However, in practice, we may not find exactly the original photo sequence but we may rearrange a collection of photos in different reasonable sequences, each of which can preserve the partial chronological order of images or image groups. It should be noticed that an image sequence which may be appropriate for one user may not be suitable for another user. One user may go shopping before going work, while another user may have a different order of these two activities/events. Therefore, we decide to exploit personal habits of a user to estimate the appropriate timespan of certain activity/event with reference to his or her personal history lifelog data. By looking into the habit of a user, i.e. the daily routines of activities that a user may perform repeatedly in many days, we may infer the most likely image sequences from different possible choices.

With the relatively discrete flow of concepts, it is challenging not only to machine learning-based models but also humans, to compare the chronological order of images. Acknowledging the nature of this shortcoming, we choose to divert our approach from processing images independently to a clustering approach. This would narrow down the discreteness to an acceptable continuity of concepts’ flow, which allows groups to be sorted efficiently at the group level.

Our key idea is that two images with high similarity regarding visual and metadata information are likely to occur in the same environment at the same timespan within a day. Thus, we first cluster images into multiple groups based on the similarity of visual and metadata information. Currently, we do not propose an approach to sort images in a single group and this will be left for future improvements. After clustering all images into separate groups, information regarding the temporal positions of those groups can be further exploited by looking back to the user’s history with the image similarity retrieval mechanism using our Bag-of-Visual-Words retrieval framework [9, 8].

We use our propose method to solve the Life Puzzle in ImageCLEFlifelog 2019. Our method achieves the best score in three criteri: Kendall’s $\mathcal{T}$, part of day, and final score. However, our method still needs further improvements to evaluate the temporal relationship of photos in each single group and to subdivide a group of photos into multiple event instances.

The content of this paper is as follows. In section 2, information about Solve my Life Puzzle challenge as well as several possible approaches are discussed. Details about our approaches in this challenge are introduced in Section 3. Experimental results together with our discussions are introduced in Section 1.

Sections 5 contains our conclusions and future works in order to enhance the performance of the proposed system.

2 Life Puzzle and Possible Approaches

To promote recent works in lifelogging problems from research groups across the world, several challenges in lifelog data retrieval had been conducted in recent years. ImageCLEFlifelog 2018 [4] which was the second edition of [2] consist of two challenge. Lifelog moment retrieval (LMRT) focuses on retrieving a specific moment in the past of a logger and daily living understanding (ADLT) aims to understand daily living in a period. From that competition, various ways to understand lifelogging data have been introduced and improved from the previous competition, with the majority of the approaches combining visual and metadata (textual, location and other information) to solve the task.

Solve my Life Puzzle challenge was first introduced in ImageCLEFlifelog 2019 [3] besides LMRT, this challenge focuses on analyzing given images along with associated metadata (e.g., biometrics, location, etc.) in a collection of images of a given day to reconstruct the correct time frame of a day and rearrange them in chronological order. To facilitate image understanding, task organizers also provides useful categories and attributes provided by Place CNN [13] and object detection results by Faster R CNN [10] trained on [6]. In this challenge, participants are given 10 queries in the test set. Each query has 25 images in a day with metadata provided. Same examples of the queries and our solutions for them are presented in Section 1.

There are several promising approaches to arrange images in reasonable sequence based on their temporal order. We can exploit the order of action/sub-activity in an event/activity to sort images in one event/activity. However, this approach may be useful when photos are taken at a small enough interval, thus photos are usually from a single event/activity.

Another approach is to exploit personal habits to estimate which timespans may be appropriate for a photo. This approach is motivated by the observation that people usually repeat to perform an activity in the same context, at the same place, and at the same timespan of the day or the same day in the week, etc. As the photos in each query of the Life Puzzle of ImageCLEFlifelog 2019 are sparsely sampled in one day (25 images in a day), we decide to follow the habit-based approach.

We adopt the Bag-of-Visual-Words retrieval framework [9] to retrieve the visually similar photos in the past history of the same lifelogger, then predict the probability for a query photo to occur in a certain timespan of a day. Our idea of applying Bag-of-Visual-Words framework to retrieve useful information in the past corresponding to a new given photo was initiated in our proposal system to retrieve similar images for reminiscence[8].

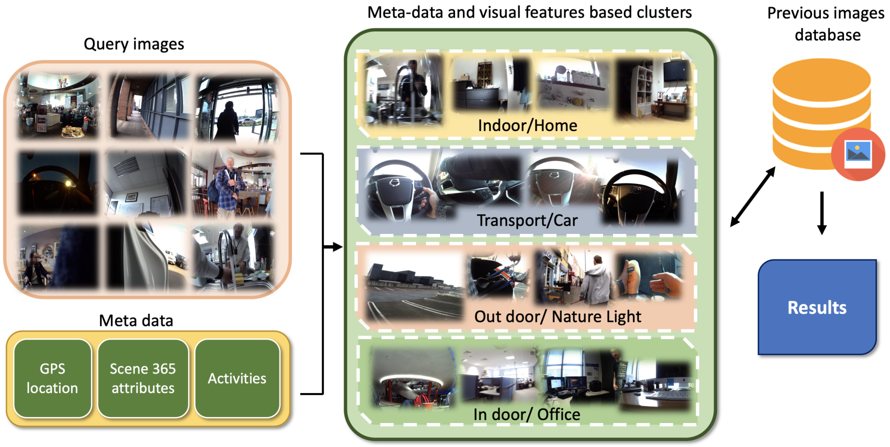

Fig. 1. Overview of our proposed system.

3 Our Proposed System

3.1 Overview of Proposed Solution

In this section, firstly we list some of the main problems to sort images in a reasonable temporal sequence as follows:

- (P1) Sorting images in one event/activity: it is necessary for the system to understand the meaning of that activity and the ordering of sub-actions in that activity. For instance, cooking should come after taking foods or drink out of a fridge and before eating breakfast.
- (P2) Sorting images in the same scene: Images should be sorted according to the visual similarity (fine grain clustering) or location information (coarse grain clustering). If a system can only use the location information, it is difficult to distinguish different scenes in the same geographical region. For example, if a user in different rooms in the same building, we cannot organize the photos simply based on GPS information.
- (P3) Sorting multiple events and actions in chronological order: the problem is how can we determine the temporal order between events and actions. In some cases, it is possible to point out the topological relationship between several pairs of events and actions, but it is extremely difficult to have a complete relationship set between every events or activities.

To solve puzzle problem for images taken at a sparse sampling rate, i.e. images are note taken at short enough time interval to belong to the same activity/event, we focus our proposed solution on the two main problems (P2) and (P3).

Figure 1 illustrates the main components of our proposed system. By extracting visual features together with metadata of each query image, a clustering system can group the input images into several clusters that have the most similarity compared to others (c.f. Section 3.2). After this step, query images are processed at the group level only. A visual-based retrieval system can look up past scenarios in the history of the same lifelogger and return possible timeframes of each image. The possible timeframe of an image group is determined by the voting scheme from the possible timeframes of each image in the group (c.f. Section 3.3).

3.2 Images Clustering

In our first phase, with all the metadata set given for each image, we cluster all of them in a small number of image groups for each photo collection (25 images) in each query. Our assumption is that photos in an image group are usually belongs to the same scene and in the same timespan. In fact, this assumption may be violated if the lifelogger stay in the same environment (office, room, shop) multiple times, i.e. event instances, in a day. However, we do not have enough visual information and given metadata to better sub-divide a group of images based on the event instances in the same environment.

In our experiments, the number of clusters ranges from 4 to 7, which are basically based on the similarities of the rough concepts. This work can be done by utilizing special concepts and attributes from Places 365 (e.g., *indoor lighting, natural light, open area, working, eating, etc.*, together with GPS location *home, work, transport*. We can also employ our visual clustering scheme [12, 11] to group images based on their visual similarity. Two examples of clustering images in the photo collection of a query are illustrated in Figure 3 and Figure 4.

3.3 Image Retrieval with Bag-of-Visual-Words Retrieval Framework

After the first phase, we already split images into multiple clusters, a.k.a. image groups, each of which is expected to belong to the same environment/context. For each image in a cluster, we use our Bag-of-Visual-Words (BoVW) retrieval framework [9] and get a ranklist of most similar images from the reference database of the past history of the same lifelogger. In our experiments, we utilize the training dataset of the Lifelog Moment Retrieval task (LMRT) of ImageCLEFlifelog 2019 as the reference database.

From the retrieved ranklist corresponding to an image, we can infer the score for each timespan when the activity in the image may be appropriate to occur. We use the voting scheme with all the scores for all possible timespans of images in a photo cluster to determine the most appropriate timespan for a cluster.

We can define a timespan to cover from 1 to 2 hours, then compare the temporal order between two clusters. Finally, we sort all clusters with reference to their temporal topological order. We also use this method with 4 pre-defined timespans to determine the part of day for each cluster.

Figure 2 illustrates our method to determine the timespan for an image cluster with respect to a reference database of the history of the same user.

For the implementation of the BoVW retrieval framework, we use Hessian-Affine detector [7] to detect keypoints in an image and Root SIFT descriptor

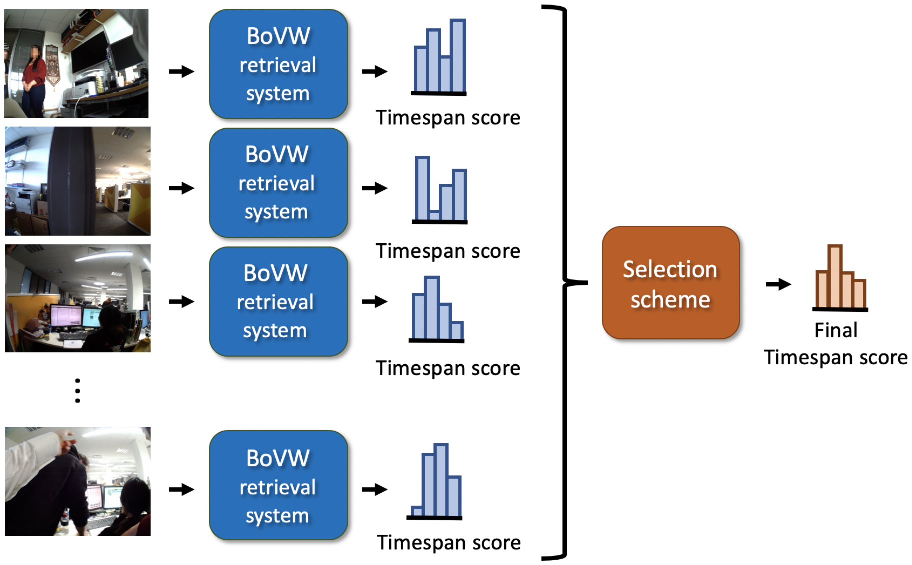

Fig. 2. Our method using Image retrieval with Bag-of-Visual-Words to determine the timespan for an image cluster.

[1] to represent the local feature in the neighbor region of a keypoint. We train a large vocabulary of one million visual words by an unsupervised clustering algorithm. The code-book is trained and stored in a server for later use to find pair of most similar images. In the second state, for each cluster, representative images are processed with the same detector and descriptor. After extraction, these features are quantized according to the pre-trained codebook using a soft-assignment strategy where each feature is assigned to the three nearest visual words. At this time, the query image is represented as a sparse BOW vector similar to those from the gallery. This vector is then independently compared to all gallery vectors using Euclidean distances. Database images having no visual words in common are irrelevant and are, therefore, filtered out quickly using the inverted index structure.

4 Results

4.1 Dataset and Task Description

The test dataset provided by organizers consists of 10 queries. There are 25 images taken at different times of day in each query. Metadata including GPS location, Scene CNN attributes, activities, etc., are provided corresponding to each image. No timestamps are given in the metadata. The challenge is how to sort all images in each query in chronological order with given information. Beside ordering, each image must be classified in 4 main categories regarding different times of the day, including morning (4h00 AM to 11h59 AM), afternoon (12h00 PM to 4h59 PM), evening (5h00 PM to 10h59 PM), night.

4.2 Examples

Several query cases with further explanations examples will be given in this section to illustrate our method.

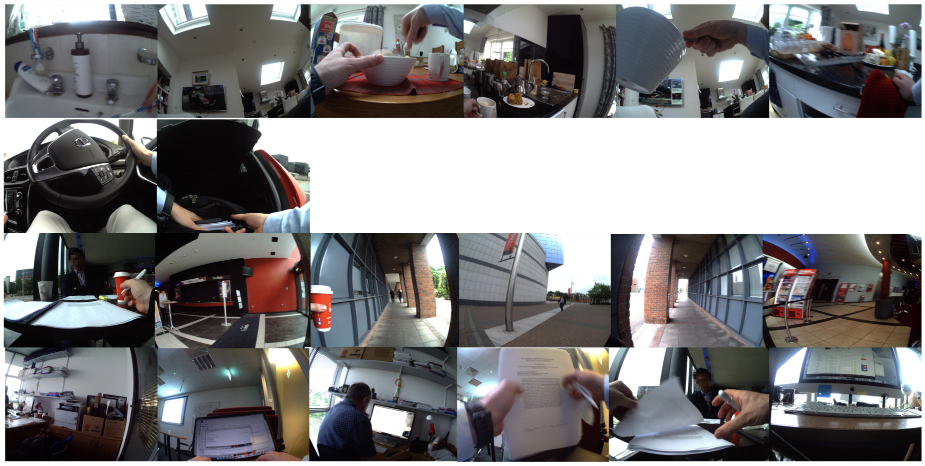

Fig. 3. Four clusters of query ID 6, each row consists of random images from each cluster.

Query ID 6: By using GPS location, Scene CNN attributes and visual features information, 25 images of the query number 6 are clustered into 4 main groups in Figure 3.

- Obviously, images were taken indoor and have a *home* tag in their GPS label are placed in the first group.
- Images taken inside a car, with a representing of *steering wheel* belongs to the second group.
- Images taken outdoor, with *natural light* belong to the third group
- Images taken indoor, with office equipment like a laptop, *papers, pen, etc.* are categorized in the last group

Query ID 9: Similar to query ID 6, 25 images of the query number 9 are clustered in to 5 main groups in Figure 4

- Images with appearances of a car are placed into the first group, due to the dominant area compared to other concepts represented in the respective images.

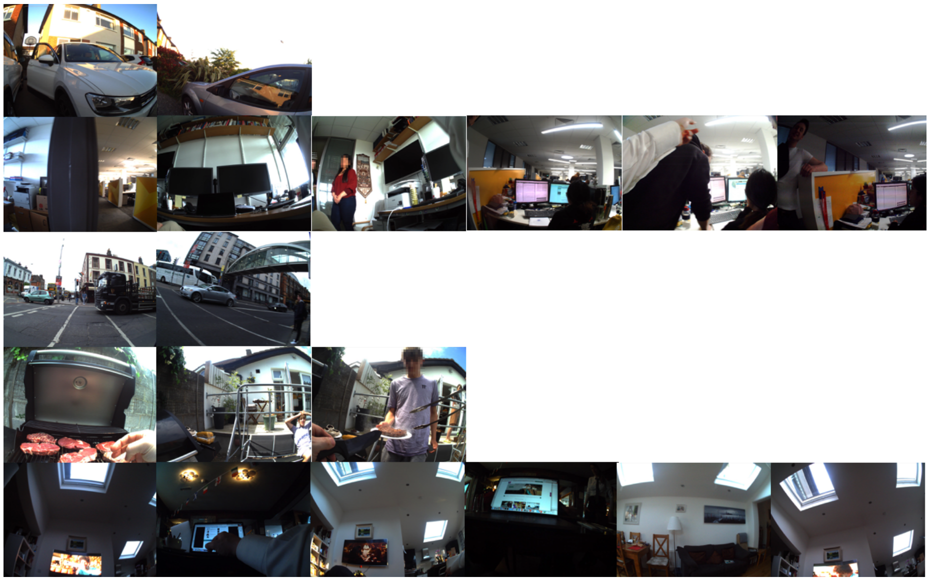

Fig. 4. Five clusters of query ID 9, each row consists of random images from each cluster.

- Images taken inside the office have a uniform distribution of detected office supplies' concepts
- Images taken on the street are characterized with high contrast and clear boundary between dark road and bright sky, with buildings.
- Some images contain a grill with its related concepts (i.e beef), are grouped into this group.
- Remaining images are taken at logger's home position.

In order to place these image groups in a timeline, different approaches have been conducted. Together with BoW images retrieval, we observe and conclude that a normal person is likely to start a day in a bathroom or have breakfast inside a kitchen. That person then needs to travel by car to his or her office by a vehicle. After walking from the car-park and having coffee with colleges, that person then stays and have various activities inside an office building. Obviously, the person is likely to finish his day at home and watch television. Regarding the grill activities, likely scenarios have been found in the database, logger usually has this kind of activity in the afternoon.

Our system's heuristic is based on the main concepts and metadata of each image to separate them into clusters. Hence, we do not deal with images in intra-cluster order yet and focus on inter-cluster order. By looking at the main concepts of a given image, we find out that it is extremely difficult to clarify the real position of each image in the timeline. For instance, with images in the second group of query 6, it is impossible to conclude the user was driving from home to the office or going back home after a working day. Similarly, we can

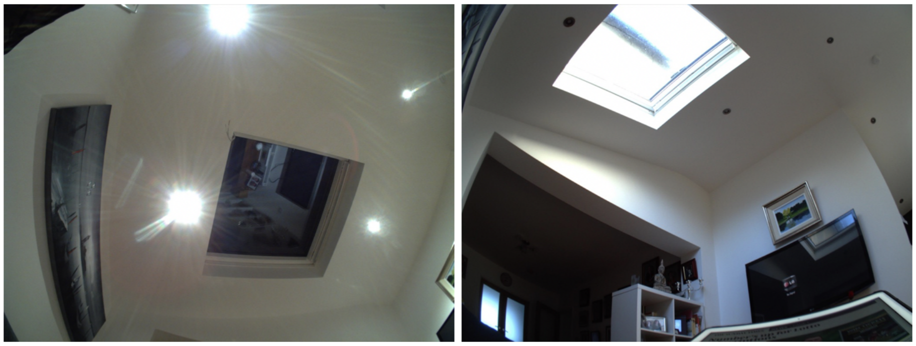

Fig. 5. Indoor images time-frame is predictable only by looking at special characteristics like lighting from windows or artificial lights.

only assume an image taken inside the logger's home may belong to the early morning or late at night. It is almost impossible to clarify unless the system can pay its attention to special details like the lighting from windows or artificial lights are turned on inside the image given (Figure 5).

4.3 Experimental Results

The official results from challenge's organizers are given in Table 1. For each team, we present their run with the highest final score. The details of submission result can be found at [3]. According to the table, our system has the highest performance among other team submissions in all three criteria: Kendall's $\mathcal{T}$, part of day, and final score.

Despite having competitive results in this challenge, the system definitely needs improvements. Currently, images belong to the same group are return in ascending filename order due to the lack of taking in specific metadata types. This leaves a potential headroom for improvements, as should more data be used for optimal retrieval instead of relying on visual information only. Future approaches can solve this remaining problem by further digest in other metadata.

Table 1. The official results of the Solve my life Puzzle challenge.

Team	Kendall's $\mathcal{T}$	Part of Day	Score
HCMUS	**0.40**	**0.70**	**0.55**
BIDAL	0.19	0.55	0.37
DAMILAB	0.02	0.47	0.25
Organiser	0.05	0.49	0.27

5 Conclusion

Re-ordering lifelog images in chronological order is a challenging problem due to the various activities that can happen in the same place across multiple time frame and the limitation of visual information. Although we cannot propose a complete solution to sort every single image in this challenge, our team proposes a method which can place clustered images in order.

Our clustering method based on the similarity of visual features and metadata of given images. Sorting inter-cluster of images can be proceeded by exploiting information in the past by using Bag-of-Visual-Words with Root SIFT features which can find the most similarity images in the image database. Our work achieved the first place winner in this year challenge with this approach.

However, there are some aspects of future improvement. The overall performance of our system can be improved if the clustering system is more robust, at that time, we can try to increase the number of clusters in each query. Besides, additional metadata utilization also proves to be a potential headroom for improvement of intra-group information retrieval in similar systems.

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. Arandjelovic, R.: Three things everyone should know to improve object retrieval. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2911–2918. CVPR '12, IEEE Computer Society, Washington, DC, USA (2012), http://dl.acm.org/citation.cfm?id=2354409.2355123
2. Dang-Nguyen, D.T., Piras, L., Riegler, M., Boato, G., Zhou, L., Gurrin, C.: Overview of ImageCLEFlifelog 2017: Lifelog Retrieval and Summarization. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
3. 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, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
4. Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of imagecleflifelog 2018: Daily living understanding and lifelog moment retrieval. In: CLEF (2018)
5. 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)

6. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014), http://arxiv.org/abs/1405.0312
7. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision **60**(1), 63–86 (Oct 2004). https://doi.org/10.1023/B:VISI.0000027790.02288.f2, https://doi.org/10.1023/B:VISI.0000027790.02288.f2
8. Nguyen, V., Le, K., Tran, M., Fjeld, M.: Nowandthen: a social network-based photo recommendation tool supporting reminiscence. In: Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia, Rovaniemi, Finland, December 12-15, 2016. pp. 159–168 (2016). https://doi.org/10.1145/3012709.3012738, https://doi.org/10.1145/3012709.3012738
9. Nguyen, V., Ngo, T.D., Tran, M., Le, D., Duong, D.A.: A combination of spatial pyramid and inverted index for large-scale image retrieval. IJMDEM **6**(2), 37–51 (2015). https://doi.org/10.4018/IJMDEM.2015040103, https://doi.org/10.4018/IJMDEM.2015040103
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. pp. 91–99. NIPS'15, MIT Press, Cambridge, MA, USA (2015), http://dl.acm.org/citation.cfm?id=2969239.2969250
11. Tran, M., Truong, T., Duy, T.D., Vo-Ho, V., Luong, Q., Nguyen, V.: Lifelog moment retrieval with visual concept fusion and text-based query expansion. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. (2018), http://ceur-ws.org/Vol-2125/paper_109.pdf
12. Truong, T., Duy, T.D., Nguyen, V., Tran, M.: Lifelogging retrieval based on semantic concepts fusion. In: Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge, LSC@ICMR 2018, Yokohama, Japan, June 11, 2018. pp. 24–29 (2018). https://doi.org/10.1145/3210539.3210545, https://doi.org/10.1145/3210539.3210545
13. 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)