Multimedia Lab @ ImageCLEF 2019 Lifelog Moment Retrieval Task

Mihai Dogariu and Bogdan Ionescu

University Politehnica of Bucharest, Romania
{mdogariu,bionescu}@imag.pub.ro

Abstract. This paper presents the participation of the Multimedia Lab to the 2019 ImageCLEF Lifelog Moment Retrieval task. Given 10 topics in natural language description, participants are expected to retrieve 50 images for each topic that best correspond to its description. Our method uses the data provided by the organizers, without adding any further annotations. We first remove severely blurred images. Then, according to a list of constraints concerning the images' metadata, we remove uninformative images. Finally, we compute a relevance score based on the detection scores provided by the organizers and select the 50 highest ranked images for submission as these should best match the search query.

Keywords: Lifelog · Information Retrieval · Visual Concepts.

1 Introduction

Wearable devices have become popular in recent years with technological advances helping in reducing their dimensions and improving their performance. Also, people have become more accustomed to interacting with gadgets, be they smart phones, smart watches, fitness bracelets, wearable cameras, etc. Moreover, lately there has been a growing exposure to multimedia content via every communication channel (e.g., TV, radio, Internet browsing, ads) up to the point where every person has had contact with or heard of wearable devices. By combining these two social trends, lifelogging emerges as a promising research field, where multimodal information is harvested and processed.

The ImageCLEF 2019 lifelog task [3] is at its 3rd edition and has gained traction over the past few years [2, 4]. It has attracted many teams in an information retrieval benchmarking competition, as part of the more general ImageCLEF 2019 campaign [10]. The purpose of the Lifelog Moment Retrieval Task is to be able to retrieve 50 images from the given dataset that correspond to a given topic (e.g., "Find the moment when u1 was using smartphone when he was walking or standing outside. To be considered relevant, u1 must be clearly using a smartphone and the location is outside."). There are 10 such topics, with different

aspects that need to be taken into consideration, such as time, location, number of objects, etc. The extracted images need to be both relevant and diverse with the official metric for the competition being the $F1@10$ measure. This metric is the harmonic mean between the precision and recall taken for the first 10 (out of 50) retrieval results for each topic.

We organize the paper as follows. In Section 2 we explore the state of the art for lifelog retrieval tasks, in Section 3 we explain our approach. Section 4 covers the experimental part and in Section 5 we draw the conclusions and discuss the results.

2 Related Work

Previous lifelogging competitions [2, 4, 3, 8, 9] have attracted numerous teams for lifelog retrieval events. Usually, teams shared a common approach to processing the input data and extract relevant information. In 2017, Molino et al. [14] won the competition with a system which filtered out blurry images and images with low color diversity and ran several CNNs on the remaining images in order to detect objects, concepts, places and persons. Afterwards, they computed a weighted relevance score for each image and selected the highest ranked ones from several clusters. In 2018, Kavallieratou et al. [11] won the competition with a system which splits the images into clusters based on location and activity and applied several CNNs to the remaining ones in order to classify them in 2,10 and 11 classes, respectively. It is worth mentioning that both of these two approaches implied the manually labeling of a part of the data.

Another relevant work is of Abadallah et al. [1] who used CNNs to extract features. Their approach also included a natural language processing component, which proves to be critical for this task. This was used to match concepts and queries, together with an LSTM. Tran et al. [16] proposed a method where they extract scene features, objects and actions. In addition textual descriptions of the images are created and then combined in an inverted index for retrieval. Tang et. al [15] and Dogariu and Ionescu [7] proposed similar techniques, where they applied a blur detection system as a first step, then extracted several types of features such as concepts, places, objects and combined them with textual knowledge.

As a general trend, most teams tried to first exclude non-informative images from the dataset and then extract several types of features, most notably objects, concepts and places. Another common aspect is that most teams needed further information for running their system, therefore they have manually annotated a part of the training data or used some sort of manual input method that would calibrate their systems. We relied on visual information and metadata parsing to solve this year's challenge, with no additional annotations and a minimum user input. The method is presented in the following section.

3 Proposed Method

Our approach focuses on excluding uninformative images as a first step and then compute a relevance score for the remaining subset of images. From previous experience, we noticed that being more strict with the criteria for excluding uninformative images leads to better results. The architecture of our system can be seen in Figure 1.

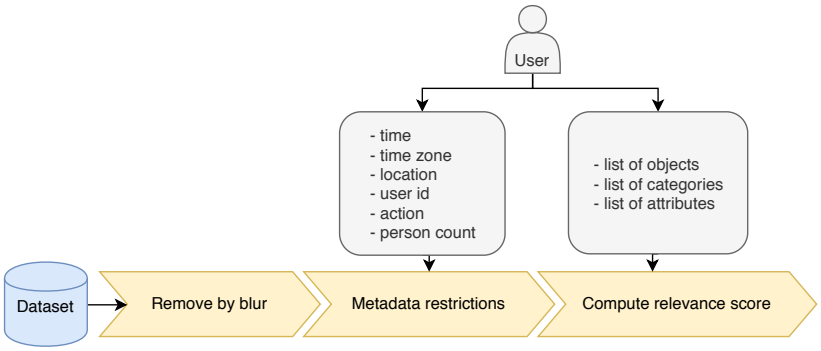

Fig. 1. Pipeline execution of proposed approach on a given query.

We start our pipeline by running a blur detection system, computing the variance of the Laplacian kernel for each image. This type of computation captures both motion blur and large homogeneous areas, when the camera might have been facing a wall or it could have been blocked by the wearer's arm. The images that do not meet a certain threshold are removed from the pipeline.

Next, the metadata of the images is checked to correspond to the restrictions imposed by the queried topic. Information about the user's id number, location, time, action, time zone are then used to remove another part of the remaining images. In some cases, this selection of metadata can suffer modifications from one topic to another, as it is described in Section 4.

At this point, the set of images on which we compute the relevance score has drastically diminished in comparison to the original dataset. Blurry images are not considered relevant in the retrieval process, but there is no metric to asses this parameter. This is a compromise, because some of the images which are part of the correct retrieval results, but have a small amount of blur, may be excluded from the processing pipeline due to our hard decision. Then, we run the remaining set of images through the relevancy score computation process.

The development dataset contained information regarding the detected attributes, categories and concepts in each image. The attributes refer to different aspects of the scenery, concepts represent the output of an object detector trained on MSCOCO [12] and the categories are the output of an image classifier trained on Imagenet [5]. We created a list of unique attributes, categories and concepts

that are present in the entire set. Then, we retained the detection confidence for each of these features under the form of sparse vectors for each image. We also kept track of the number of detections of each object in every image, since each object could trigger multiple detections in the same image. For the final part of our algorithm, we manually select several attributes, categories and concepts which are relevant to the queried topic. These can mean that they must either be present or not present in the image's detection results, depending on the topic. More details on this are given in Section 4.

Having all the needed information available, we proceed to computing the relevance score as the sum of the confidences of features that need to be detected in the image. Mathematically, the score S for a single image is expressed as follows:

$$S = \sum_{i=1}^{|A|} \{s(a_i)|a_i \in L_A\} + \sum_{i=1}^{|C|} \{s(c_i)|c_i \in L_C\} + \sum_{i=1}^{|O|} \sum_{j=1}^{|O_i|} \{s(o_{i,j})|o_i \in L_O\}, \quad (1)$$

where A is the set of all attributes, a_i is one attribute from this set, L_A is the subset of attributes that we selected as being relevant for the query and they must be present in the image's detection results and $s(a_i)$ is the confidence score of the respective attribute, a_i. Similar reasoning is applied for the categories set, C. We denoted the concepts set with O, since these concepts are, in fact, objects from the MSCOCO dataset. As each object can appear several times in a single image, we denote the subset of its detections with $|O_j|$. $s(o_{i,j})$ refers to the confidence of the j^{th} detection of object of index i. As opposed to [7], we do not use a weighted sum because the weights have to be manually tuned for each individual query and it would stray too much from the idea of automatic processing.

This score is computed for each image and we rank them in descending order according to it. From previous experience, we saw that many events targeted by topics have a low level of recurrence. In other words, they are isolated events which occur only once in the dataset. Therefore, we decided to improve precision, rather than cluster recall, so we did not apply any diversification method. In the end, we submitted the 50 best ranked images per topic, according to the relevance score.

We note that the user has to manually select the parameters for both the metadata restrictions and the list of items that drive the relevance score and this is the only manual input required from the user. As far as we know, there has yet to be developed a clear method on how this parameter tuning process could be completely automatized.

4 Experiments

The development dataset and the assigned topics proved to be challenging, as in previous editions. There was a plethora of data concerning each image that

had to be stored under a homogeneous format. Moreover, since there is no fixed template for the conditions imposed by different queries, it is important to have as much information as possible about individual images. Therefore, we stored all the available metadata for all images, even if we only used just a part of it in the retrieval process. To give an idea regarding the magnitude of this effort, we stored 31 different data fields for each of the 82k images in the dataset. We note that 4 of the 31 data fields represented feature vectors of lengths 77, 77, 99 and 302, respectively.

Our approach involved a common part to all 10 queries, namely the removal of images with a high amount of blur. As mentioned in Section 3, we computed the variance of the Laplacian for each image. We imposed a threshold of 90 for the blur filter which is more aggressive than in previous works as we observed that many images from the final list were previously still suffering from motion blur. The dataset is thus reduced from about 82k images down to roughly 35k images. This is a major reduction in the number of candidate images for the next part of the pipeline.

In the metadata restrictions that we imposed on the images we used 6 types of parameters: the user's id, the action that was being performed, the local time, the location, the time zone and the person count. We used different combinations of these metadata, with the notable exception of user id and action, which were present in all combinations. We relied on the detection result of the object detector to account for the number of persons in the scene. In Table 1 we summarized the types of information that were used for each topic, marking with an 'X' the type of parameter that was used for the respective topic. A complete description of the topics can be found in the competition's overview paper [3].

Table 1. Metadata combinations used for each topic - × indicates that the respective type of information was used

Topic number	*User id*	*Activity*	*Location*	*Time*	*Time zone*	*Person count*
1	×	×	×	×	-	-
2	×	×	×	×	-	-
3	×	×	×	-	-	-
4	×	×	×	-	-	-
5	×	×	×	-	-	-
6	×	×	×	×	-	-
7	×	×	×	×	-	×
8	×	×	-	-	-	-
9	×	×	-	-	-	×
10	×	×	-	-	×	×

It is important to have a detailed discussion on how and why these parameters were chosen, depending on the queried topic's description. All but one topic, T4, required retrieving images corresponding to u1. However, during submission of the proposed list of images for user 2, on T4, we encountered a problem from

the submission platform, which rendered all our proposed images corresponding to u2 as erroneous. Therefore, we eliminated all images from user 2 from our submission. For the activity field, only topic T2 asked for the user to be driving, therefore we imposed the restriction that images should have the "transportation" activity. For all other images we imposed that the user should have any other activity than "transportation".

The location where the images were taken was also helpful as we had to identify different sequences where the user was at home, in a cafe or leaving from work. We did not impose drastic restrictions regarding the metadata, since we did not want to eliminate potentially relevant images. We also tried to find the moments when the user was in a toy shop for T1 based on the location, but noticed that the location was not synchronized with the respective images, having a gap of about 1 hour in between the images that were taken inside the shop and the images that were annotated as being inside a toy shop. Therefore, location was not decisive for this topic, as we expected.

Time constraints were imposed only regarding the working hours of regular shops, for T1, the user's working hours, for T2 and T7 and the time interval imposed by T6's topic description. We also used the time zone of the images in order to locate the set of images when the user travelled to China. Lastly, the person count from the concept/object detector was used in topics which specifically asked for a certain number of people to be present in the image.

Following this step, we compiled a list of categories, concepts and attributes that might fit each topic, individually. One case in which this technique was somewhat successful is for T1, when we were asked to find the moments when the user was looking at various toys, such as electronic trains, model kits and board games. Here, we searched for attributes such as {"playing", "shopping", "gaming", "plastic", "cluttered", "supermarket"}, all of which could be related to a toy shop. These helped us create the attributes list, L_A from eq. 1. For the categories list, L_C, we selected {"store", "shop", "toy", "train", "arcade"}. We also searched for objects such as {"board", "game", "train", "toy", "model", "kit", "bus"} in the image. These represented the objects list, L_O. We applied similar reasoning for the rest of the topics. The length and variety of items that could fit in either of the 3 aforementioned lists changed from one topic to another. For some of them it worked well, whereas for others it gave us disappointing results. The full set of attributes, categories and objects that we used for each topic can be seen in Table 2.

A somewhat different approach was for topic T2, "Driving home", where the participants were asked to retrieve images when u1 was driving home from the office. Any other departure or arrival point than the ones mentioned in the description render the image irrelevant. Here, we considered that this event can happen at most once each day. Then, we took all the images from the afternoon (time interval between 16 and 20 o'clock) and kept only the ones that had the "transport" label for their activity. This should reduce the set of images to only the ones when the user was driving. Afterwards, we checked to see if there was any pause between successive images, when the user was not driving anymore.

Since we noticed that the "transport" action is continuous throughout an entire car drive interval, having a pause in this interval would mean that the user had an intermediate stop on his way home and would remove the image from the list.

Table 2. Attributes, categories and objects lists used for each topic

Topic	Attributes, categories and objects lists
1	L_A = ['playing', 'shopping', 'gaming', 'plastic', 'cluttered', 'supermarket'] L_C = ['store', 'shop', 'toy', 'train', 'arcade'] L_O = ['board', 'game', 'train', 'toy', 'model', 'kit', 'bus']
2	L_A = ['driving', 'glass', 'metal', 'matte', 'glossy', 'transporting'] L_C = ['car interior', 'bus interior', 'cockpit'] L_O = ['car', 'bus']
3	L_A = ['indoor', 'eating', 'plastic', 'cluttered', 'shopping', 'vegetation'] L_C = ['kitchen', 'shop', 'ice', 'living room'] L_O = ['pizza', 'bottle', 'broccoli', 'refrigerator', 'sandwich', 'cup']
4	L_A = ['indoor', 'cluttered space', 'enclosed area', 'sports', 'spectating'] L_C = ['living room', 'soccer field', 'soccer', 'football', 'television'] L_O = ['tv', 'sports ball', 'remote']
5	L_A = ['indoor', 'socializing'] L_C = ['coffee shop', 'cafeteria', 'bar', 'restaurant', 'lobby'] L_O = ['cup', 'person']
6	L_A = ['indoor', 'eating', 'cluttered space', 'enclosed area', 'plastic'] L_C = ['kitchen', 'living room', 'dining room', 'food', 'pizzeria', 'picnic'] L_O = ['pizza', 'bottle', 'sandwich', 'cup', 'dining table', 'cake', 'toaster']
7	L_A = ['indoor', 'socializing', 'cloth'] L_C = ['coffee shop', 'cafeteria', 'bar', 'restaurant', 'lobby'] L_O = ['cup', 'person']
8	L_A = ['natural', 'pavement', 'concrete', 'vegetation', 'trees', 'sunny'] L_C = ['outdoor', 'phone booth', 'park', 'street', 'garden'] L_O = ['cell phone', 'cell phone', 'cell phone', 'car', 'bus', 'traffic light']
9	L_A = ['glass', 'glossy', 'natural light', 'indoor', 'manmade'] L_C = ['indoor'] L_O = ['person']
10	L_A = ['business', 'indoor lighting', 'man-made', 'paper', 'research'] L_C = ['indoor', 'restaurant', 'conference', 'classroom', 'lobby'] L_O = ['person', 'suitcase', 'bottle', 'chair', 'dining table']

The official results of our run can be seen in Table 3. The obtained precision rate was lower than expected. Moreover, high contrast to the cluster recall on several topics also affected the overall F1 measure.

We submitted 2 runs, but the second one followed the same algorithm, with the only exception being that it excluded pictures taken by the user with his

Table 3. Official results of our submitted run.

Topic number	*P@10*	*CR@10*	*F1@10*
1	0.2	0.5	0.285
2	0.3	0.047	0.082
3	0.1	0.055	0.071
4	0	0	0
5	0.7	0.22	0.33
6	0	0	0
7	0.1	1	0.181
8	0	0	0
9	0	0	0
10	0.3	0.33	0.31
Mean	**0.17**	**0.215**	**0.127**

mobile phone. However, it obtained the same exact result so we decided to only present the results of the first run.

We also present the final state of the learderboard, at the end of the competition in Table 4. Our approach ranked 8th out of 10 teams. The entries field refers to the number of times that each team tried to submit a run. This accounts for both valid and wrong submissions. Therefore, it is not to be confused with the number of different runs.

Table 4. Leaderboard

Position	Team name	F1@10	Entries
1	HCMUS	0.61	4
2	ZJUT	0.44	8
3	NICT	0.367	3
4	Baseline	0.289	
5	ATS	0.255	20
6	DCU	0.238	5
7	Regim_Lab	0.188	10
8	**Multimedia Lab (ours)**	**0.127**	**5**
9	TU Chemnitz	0.117	16
10	University of Aveiro	0.057	7

This year there were a total of 10 teams competing in the LMRT task, the most that have been recorded in Image CLEF Lifelog competitions since they began. We can see that the leader stands far from the rest, while the rest of the ranking remains quite balanced. This proves that lifelogging is gaining traction and draws the attention of more and more research teams in a highly complex challenge.

5 Discussion

The algorithm that we proposed is composed of a selection of the images based on the amount of blur, the metadata that is associated to them and then computing a relevance score in accordance with several manually built lists of items that best describe each particular topic query. We took into consideration an automatic selection of these parameters, but, as reported in [6], this is not a trivial task. We also took into consideration using a word2vec model [13], but not having enough documents relevant for our task made it difficult to extract relevant meanings for the words inside the topics' descriptions. Therefore, being able to automatically go from a natural language description of the topic to a list of accurately defined terms which best describe the relevant images still remains an open problem.

Another important aspect is that given the large variety of aspects that are searched for in the set of images, it is difficult to propose a unique system that would solve all 10 queries. Several tuning mechanisms are in order to help the overall architecture adapt to particular tasks. Additionally, training neural nets on the available data, without supplementary annotations, could lead to overfitting, since the provided groundtruth is very scarce in comparison to what neural nets need for being robust enough.

Once the groundtruth data will be available we plan to run ablation studies to figure out the way each part of our system had an impact on the overall performance, both positive and negative. In conclusion, lifelog moment retrieval remains a challenging task in which it is crucial to understand the provided data and the limitations of current state of the art. Many efforts have been made in this direction due to the Image CLEF Lifelog campaigns, encouraging researchers to take part in this benchmarking competition.

Acknowledgement

This work was supported by the Ministry of Innovation and Research, UEFISCDI, project SPIA-VA, agreement 2SOL/2017, grant PN-III-P2-2.1-SOL-2016-02-0002.

References

1. Abdallah, F.B., Feki, G., Ezzarka, M., Ammar, A.B., Amar, C.B.: Regim lab team at imageclef lifelog moment retrieval task 2018 **2125** (September 10-14 2018)
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, vol. 1866. CEUR-WS.org <http://ceur-ws.org/Vol-1866/>, Dublin, Ireland (September 11-14 2017)
3. Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C.: Overview of ImageCLEFlifelog 2019: Solve my life puzzle and Lifelog Moment Retrieval. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org <http://ceur-ws.org/Vol-2380/>, 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: CLEF2018 Working Notes. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org <http://ceur-ws.org/Vol-2125/>, Avignon, France (September 10-14 2018)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Dogariu, M., Ionescu, B.: A textual filtering of hog-based hierarchical clustering of lifelog data. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. vol. 1866. CEUR-WS.org
7. Dogariu, M., Ionescu, B.: Multimedia lab @ imageclef 2018 lifelog moment retrieval task. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. vol. 2125. Avignon, France (September 10-14 2018)
8. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Albatal, R.: Overview of ntcir-12 lifelog task. In: NTCIR. Pisa, Italy (July 2016)
9. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Gupta, R., Albatal, R., Nguyen, D., Tien, D.: Overview of ntcir-13 lifelog-2 task. NTCIR (2017)
10. 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), vol. 2380. LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
11. Kavallieratou, E., del Blanco, C.R., Cuevas, C., García, N.: Retrieving events in life logging **2125** (September 10-14 2018)
12. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
14. Molino, A.G.d., Mandal, B., Lin, J., Lim, J.H., Subbaraju, V., Chandrasekhar, V.: Vc-i2r@ imageclef2017: Ensemble of deep learned features for lifelog video summarization **1866** (September 11-14 2017)
15. 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 **2125** (September 10-14 2018)
16. Tran, M.T., Dinh-Duy, T., Truong, T.D., Vo-Ho, V.K., Luong, Q.A., Nguyen, V.T.: Lifelog moment retrieval with visual concept fusion and text-based query expansion **2125** (September 10-14 2018)