LIG-MRIM at NTCIR-12 Lifelog Semantic Access Task

Bahjat Safadi[1,2,*], Philippe Mulhem[1,2,*], Georges Quénot[1,2,*], and Jean-Pierre Chevallet[1,2,*]

[1]Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
[2]CNRS, LIG, F-38000 Grenoble, France
[*]Firstname.Lastname@imag.fr

ABSTRACT

This paper describes the participation of the LIG-MRIM research team to the Lifelog Semantic Access sub-task of the NTCIR-12 (2016). Our approach mainly relies on mapping the query terms to visual concepts computed on the Lifelogs images according to two separated learning schemes. A post-processing is then performed if the topic is related to temporal, location or activity information associated with the images. The results obtained are promising for a first participation to such a task, with event-based MAP above 29% and an event-based nDCG value close to 39%.

Team Name

MRIM

Subtasks

Lifelog Semantic Access Task

Keywords

Visual concepts, deep learning, temporal description, ImageNet, TRECVid

1. INTRODUCTION

The MRIM team of the Laboratory of Informatics in Grenoble (LIG), in France, participated to the pilot Lifelog Semantic Access Task (LSAT)[1], [3].

According to the data provided and typical queries, we considered two facets of the Lifelog images: visual and temporal. We processed each image of the corpus in a way to extract visual concepts according to two different vocabularies, namely ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [7] and TRECVid [6] concepts. The images are also characterized by one or several of temporal concepts. These visual and temporal concepts serve as a basis for the retrieval: a first step focuses on visual concepts (i.e. "what do we see?") and then we filter the results by temporal aspects when needed (i.e. "when does it take place?"). All the runs submitted are "automatic" according to the official definition of the Lifelog Semantic Access Task, as there was no user involvement in the search beyond specifying the query.

The rest of the paper is organized as follows. We first present a short overview of the Lifelog task and the provided data [1] in Section 2. Then, we focus on the image indexing using a framework based on Deep Learning models and on MSVM classifiers in section 3. Section 4 focuses on the temporal aspects of the images, by describing a simple binary mapping into predefined time slots, such as "early morning". Section 5 depicts how the data from the *semantic* tags (e.g. "location" and "activity"), automatically assigned for each frame, are integrated in our description. Section 6 explains how, based on these elements, we provide a way to process queries (or *topics*). In practice, we relied on manual expressions of queries that simulates an automatic mapping into visual and temporal concepts. The official results obtained are presented and commented in Section 7, before concluding in Section 8.

2. TASK OVERVIEW

The goal of the LSAT task is to retrieve a number of specific moments in a lifelogger's life. Moments are semantic events, or activities that happened throughout one or several days. NTCIR-Lifelog data consist of anonymised (faces and names removed) lifelogs gathered by a number of individuals over an extended period of time. There are two data sets for NTCIR-12 Lifelog pilot task:

- Dry Run data set consisting of one day of data from two lifeloggers. This will allow for participants to prototype their retrieval systems and submit test results.
- Full NTCIR-12 Lifelog data set. As described above, a 100 day data set from a number of lifeloggers. This is the data set that we will use for the evaluation.

Each of the two NTCIR data sets contains:

- Images taken automatically by the lifelog device;
- Visual Concepts (automatically extracted visual concepts with varying rates of accuracy);
- Semantic Content (semantic locations, semantic activities) based from sensor readings on mobile devices.

3. VISUAL INDEXING

The visual indexing of the lifelog images is present in Figure 1. It it composed on two main parts:

- In a first step each image is processed with three different Deep Convolutional Neural Network models using the *caffe* framework [4], namely the AlexNet network [5], the VGG network [11] and the GoogLeNet

[1] http://ntcir-lifelog.computing.dcu.ie/

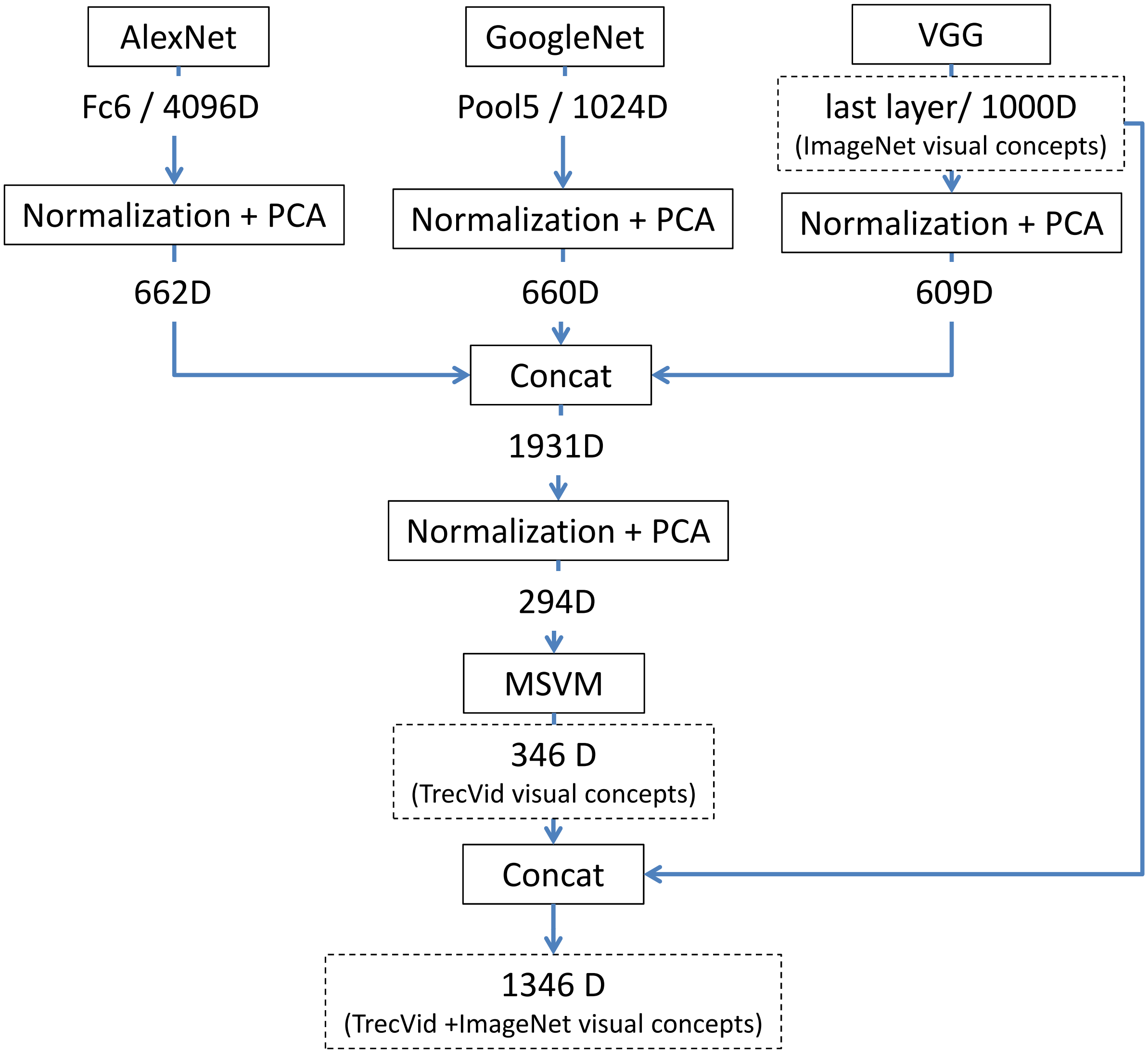

Figure 1: MRIM Lifelog Visual indexing

network [12]. Each of these networks have been learned on the ImageNet corpus. To take advantage of several categories of features, we consider the last output layer of VGG (i.e., 1000 visual concepts of ImageNet), the layer fc6 from the AlexNet (non visual concept features, just above the feature convolutional layers), and the pool5 layer from GoogLeNet (non visual layer below the final layer). The idea is that the different kinds of features extracted may better represent different visual facets of images. Moreover, as the output from VGG describes visual concept, such representation will be used to link the topics' terms to ImageNet concepts (dotted box);

- In a second step we use another set of terms that are able to describe the visual content of images. This set comes from the well-known TRECVid evaluation campaign, and is composed of 346 concepts. This set does not overlap with the ImageNet concepts. To learn the models for such concepts, we made use of the Multiple-SVMs (MSVM) approach [9], mainly we used the accelerated version of the MSVM [10], learned on the TRECVid 2013 data. The output vectors from the three considered networks, were merged and used as an input descriptor to the MSVM. However, these vectors were optimized separately using the power-law and PCA approach [8], as well as the same approach was applied to optimize the merged descriptor to produce the final descriptor, which has a 294 dimensions. For each concept of TRECVid, we trained a MSVM model using the merged descriptor of 294 dimensions. This results in 346 models. For efficiency, these models were merged together in one global model following the FMSVM [10] approach. For the lifelog images, we used the global model to predict the existence of the 346 concepts in the images. These predicted values are used as linkage to topics terms when needed (dotted box).

Given a topic "query", we link the terms of the topic manually to the set of ImageNet and TRECVid concepts. For the visual representation of the topic, we merge the scores of the linked concepts (from both sets the ImageNet and TRECVid). Therefore, each image is scored according to the selected concepts that fit with the topic. This process is currently achieved manually, but we believe that in most of the cases such mapping may be automatic.

4. TEMPORAL INDEXING

In addition to storing the provided date/time of each frame, the temporal indexing of images is a very simple one: we named several hours of the day according to table 1 (top), that do not take into account the day of the week. Such table allows overlapping of time slots, as these concepts are quite fuzzy and culturally dependent. Others concepts depend on the day of the week, as they are more related to working events, as described in table 1 (bottom). These temporal concepts are binary, and are used to tag each lifelog image of the corpus.

Table 1: Temporal indexing terms

Time slot	Days	name
21:00 PM - 5:00 AM	All	night
5:00 AM - 7:15 AM	All	early morning, breakfast
7:30 PM - 11:30 AM	All	morning
11:30 AM - 2:00 PM	All	lunch
2:00 PM - 17:30 PM	All	afternoon
17:30 PM - 20:00 PM	All	early evening
20:00 PM - 23:00 PM	All	late evening
7:30 AM - 9:00 AM	Mon-Fri	trip from home to work
18:15 PM - 18:45 PM	Mon-Fri	trip from work

5. LOG INDEXING

We also integrated the location and activity fields (as character strings) of each frame in the lifelog to index the images.

6. QUERY PROCESSING

The query processing is manual and based on two steps that consider in sequence the elements described above.

- The first step relies on the visual concepts that are detected on the lifelog images, using ImageNet and TRECVid concepts as indexing concepts. More precisely, we begin by checking from the topics the visual terms from TRECVid and ImageNet concepts lists. A non-weighted linear combination of scores is then processed when more than one visual concept is selected, to produce a visual score for each image. Furthermore, images are ranked according to their visual scores.
- The second step is built as a filter among the result lists obtained at the end of the first step: if any topic's term matches any temporal, location or activity concept, then it is used to filter the result. If no term is found then no filtering is processed.

Our approach does emphasize the concepts aspects of queries first, and focuses in a second step on the other information (temporal, activity, etc.). Theoretically however, it is like ANDing both aspects for most of the case (i.e. all topics but 2, that need a complex integration of temporal aspects, see below).

7. OFFICIAL RESULTS DISCUSSION

7.1 Overview

The official evaluation of the LSAT task is based on two levels: image-level and event-level. For the image-level results, the relevance of each image to the topic in question is checked. For the event-level results, every image included in a submission is mapped to the event that it belongs to, and the results are then calculated at the event-level. The evaluation measures are classical for IR systems: Normalized Discounted Cumulative Gain (nDCG) and Mean Average Precision (mAP). The results according to the two levels and the two evaluation measures are presented in table 2. The results obtained are very good, for a first participation in a Lifelog retrieval campaign, especially for the event-based mAP and nDCG evaluations.

What we conclude from these general results is that, as expected, our event-based measures are higher than image-based one, as event-based measure tend to favor precision instead of recall. Having lower values, for image-based results, may be due to: i) the time/location/activity-based filtering is probably too rigid, or ii) there exist some kind of instability in the visual indexing.

Table 2: MRIM results official evaluation

level	nDCG	mAP
Event	0.3896	0.2940
Image	0.2455	0.1667

7.2 Details

Among queries generated from the 48 initial topics provided by the task organizers, two of them (corresponding to topics 009 and 048) led to an empty query. From the remaining 46 topics, we get the following statistics

- for the visual concepts: 29 include TRECVid concepts only, 13 include ImageNet concepts only, and 3 use concepts from both TRECVid and ImageNet;
- only two queries use temporal concepts;
- two queries involve more complex integration of the time aspect. The query from topic 35 use explicit dates: we assume first that the log begins in the city of the logger. Then we detect when the user is at the airport, and we filter the day in between before filtering the initial set of images. The query from topic 32 is related to having a journey after being at the airport; so we first select moments when the user is at the airport, and we focus on the time frames that are posterior to the stay at the airport;
- 30 queries from the topics include an explicit usage of the location tag, assuming an explicit knowledge of the life-loggers;
- 21 queries from the topics make use of the activity tags, mainly to find transportation events (transport, cycling, walking). Additional 13 queries use explicit negations of any activity (meaning images that are not associated with any activity). Such negation indicates that the user is expected to be static (for instance when drinking with friends).

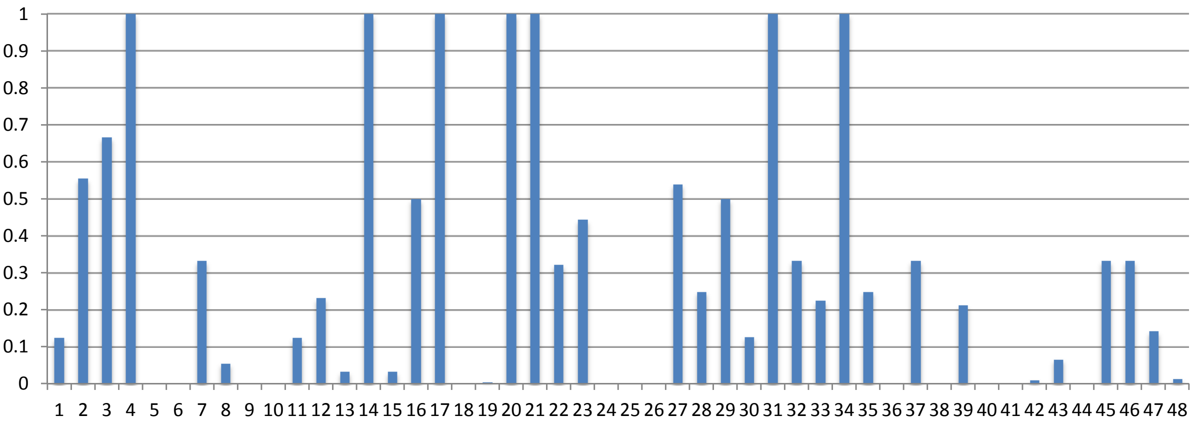

Figure 2: Event-based AP official result per topic.

We discuss now the results query by query, by focusing on the event-based Average Precision results as presented in Figure 2. We limit our comments on this evaluation measure as the results are comparable to the other official measures (event nDCG, and image map and nDCG). We see in Figure 2 that seven results (topics 4, 14, 17, 20, 21, 31, 34) achieve an AP of 1.0 . In the related queries all but one based on MSVM visual concepts, the remaining one uses VGG visual concepts. For the 12 null AP results with visual concepts, 8 of then use MSVM only visual concepts, 3 of them use VGG only visual concepts, and 1 uses both concepts from MVSM and VGG.

We have not been able to determine a link between the presence/absence of location/activity and the quality of the results. The impact of the temporal features are also not obvious, as queries containing temporal criteria have respectively APs of 0.0, 1.0, 0.33 and 0.25. In fact, it is clear that our initial choice of putting the priority first to the visual elements, and then only to post-filter the initial results using the temporal/location/activity features does not provide way to analyse, exclusively, temporal/location/activity features.

After carefully checking the results obtained, we found out that few (three) of the queries we generated are incorrect, especially according to the spelling of some locations. We will have to rerun the correct queries to see if it impacts positively our overall results.

8. CONCLUSION

We proposed a way to retrieve events in a lifelog data stream. Even if we fit the definition of "automatic" runs for the task, we did generate manually the queries from the topics. According to the protocol we used to express the queries, we believe that "true" automatic processes should be able to achieve similar results, at least for the visual aspects of the topics.

In the future we will focus on such automatic mapping and routing into conceptual/temporal concepts. Word embedding approaches like [2] may be relevant in our case. Other research questions are related to the way we process queries: our approach is like ANDing the visual concept aspects and the other aspects, but more *fuzzy* fusions of these aspects may be more effective. It is clear also that some complex queries (according to the visual aspects, or the temporal aspects) might also be studied separately to be able to be properly tackled.

Acknowledgements

This work was partly realized as part of the CHIST-ERA Camomile Project funded by ANR, French national research agency. This work has been partly carried out in the context of the Guimuteic project funded by Fonds Européen de Développement Régional (FEDER) of région Auvergne Rhône-Alpes.

9. REFERENCES

[1] G. Cathal, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal. Ntcir lifelog: The first test collection for lifelog research. In *SIGIR*, pages 1–1. ACM, 2016.

[2] Y. Goldberg and O. Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722, 2014.

[3] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal. Overview of ntcir lifelog task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-12*, pages –, 2016.

[4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

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

[6] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Quénot, and R. Ordelman. Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, April 2015.

[8] B. Safadi, N. Derbas, and G. Quénot. Descriptor optimization for multimedia indexing and retrieval. *Multimedia Tools and Applications*, 74(4):1267–1290, 2015.

[9] B. Safadi and G. Quénot. Adaptivity, personalization and fusion of heterogeneous information. In *CBMI*, pages 88–91, 2010.

[10] B. Safadi and G. Quénot. A factorized model for multiple svm and multi-label classification for large scale multimedia indexing. In *Content-Based Multimedia Indexing (CBMI), 2015 13th International Workshop on*, pages 1–6. IEEE, 2015.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.