Lifelog Semantic Annotation using deep visual features and metadata-derived descriptors

Bahjat Safadi, Philippe Mulhem, Georges Quénot, Jean-Pierre Chevallet

Lifelog Semantic Annotation using Deep Visual Features and Metadata-Derived Descriptors

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 a method for querying lifelog data from visual content and from metadata associated with the recorded images. Our approach mainly relies on mapping the query terms to visual concepts computed on the Lifelogs images according to two separated learning schemes based on use of deep visual features. A post-processing is then performed if the topic is related to time, location or activity information associated with the images. This work was evaluated in the context of the Lifelog Semantic Access sub-task of the NTCIR-12 (2016). The results obtained are promising for a first participation to such a task, with an event-based MAP above 29% and an event-based nDCG value close to 39%.**

I. Introduction

The concept of Lifelogging has been gathering important attention by reasearchers in Multimedia Indexing In recent years. Lifelogging represents a phenomenon whereby individuals can digitally record their own daily lives, in varying amounts of detail and for a variety of purposes. In a sense it represents a comprehensive black-box of a person's life activities and offers great potential to mine or infer valuable knowledge about life activities, given the availability of appropriate software [1], [4]. Moreover, recent advance of wearable devices and other sensors started to enable a wider range of people to perform data-oriented via lifelogging, and makes the concept more and more popular. Thus, massive multimedia archives are continuously produced, as every moment of life-experiance is captuared and recorded to represent a lifelog. Such a lifelog needs to be indexed, organised and searchable to be valuable to the lifelogger.

Typically, each lifelogger produces over 1000 images per day (every 30 seconds in average) and some already did it for many years. The motivation for recording them varies among lifeloggers, but they all need efficient tools for exploiting the recorded data. A typical application would be to answer questions like: "when did I meet X in place Y" or "find the instances when I had this meal for dinner" or many other similar ones. While recording already reaches millions of images for many lifeloggers, almost no tool already exist for such applications. However, the dedicated devices used for recording the images are now able to record associated information like date and time, location (e.g. from GPS) and sometimes even activities like walking (from MEMS accelerometers and/or gyroscopes).

In order to encourage research on automatic indexing and retrieval in lifelog data, the first test collection for lifelog research was released [15] as part of the NTCIR-12 (2016) project [1]. This test collection are highly individual and multimodal when compared to conventional test collections. Two pilot tasks were released:

1) Lifelog Semantic Access Task (LSAT) to explore search and retrieval from lifelogs, and
2) Lifelog Insight Task (LIT) to explore knowledge mining and visualisation of lifelogs.

The two tasks comes with set of queries and test data, and no training data of lifelogging is provided. We evaluated our approach in the context of the first one (LSAT).

According to the data provided and typical queries, we considered two facets of the Lifelog images: the content-based (visual) one and the metadata-based (textual) one, including time, location and activity information. We processed each image of the corpus in a way to extract visual concepts according to two different vocabularies, namely the 1000 ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2012) [8] and the 346 TRECVid [7] concepts. The images are also characterized by one or several of temporal concepts. These visual and metadata-based 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 (and/or location/activity) aspects when needed (e.g. "when does it take place?").

In this paper, we investigates the following questions: i) to which extent can visual concepts contribute relevant information when searching lifelog images with natural language queries? ii) can the narrative information of a query provided alongside the text query be mapped to their corresponding visual concepts? This mapping is currently done manually, but automating this process is in the perspective of the paper. iii) How can we use the temporal information, generated automatically by the device while taking images, help in

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

978-1-4673-8695-1/16/$31.00 2016 IEEE

filtering the results of visual indexing? While addressing those important questions, the main contributions of this paper are summarized as follows:

- We propose a framework which integrates semantic visual indexing with text queries for lifelog image retrieval.
- We demonstrate the effectiveness of the proposed framework in narrowing the semantic and intention gaps (to better answer the user query) in large-scale lifelog image retrieval.

The rest of the paper is organized as follows. We first present a short overview of the Lifelog task and the provided data in Section II. Then, we focus on the image indexing using a framework based on Deep Learning models and on MSVM classifiers in section III-A. Section III-B focuses on the temporal aspects of the images, by describing a simple binary mapping into predefined time slots, such as "early morning". Section III-C depicts how the data from the *semantic* tags (e.g. "location" and "activity"), automatically assigned for each frame, are integrated in our description. Section III-D 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 IV, before concluding in Section V.

II. Task overview

In this paper, we address the task a user faces when searching specific moments in a lifelogger's life. Such moments are semantic events, or activities that happened throughout one or several days. The search is initiated by a query formulated by a user using free form text (words, groups of words or sentences) and augmented by a Narrative description of the visual content, which may include moment time, locations or activities, etc. Therefore, we focus on the LSAT task at the NTCIR-Lifelog [3], [15], that addresses the same goal of this paper. The provided 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.

The LSAT task has 48 queries that fits the goals of this paper, in which users provided queries is constituted of two textual parts: the searched query for visual content, the other for the Narrative information depicting the searched moments (temporal). Here, we give two examples of such a query:

1) **id:** 014; **type:** precision; **uid:** u1; **title:** The Church
   description: Find the moment(s) when I was inside the main hall of a Church.
   narrative: To be considered relevant, the moments much show the user inside the main hall of a small church. Being inside the church is recognisable by the presence of a cross on the wall. Standing outside the hall of the church is not relevant. Moments showing external views of a church are also not considered relevant.
2) **id:** 031; **type:** precision; **uid:** user2; **title:** Bus to the Airport
   description: Find the moment(s) in which I was taking a bus to an airport.
   narrative: To be considered relevant, the user must be riding a bus, the final destination of which is an airport. The user must be going to the airport and not coming from the airport. It does not matter in which country the moment takes place.

III. Proposed framework

The proposed framework is shown in figure 1. It is composed of two major parts depending on whether the processing is performed at query time (on-line) or in advance in the background (off-line). Based on a lifelog image collection (e.g. images with their temporal information) and a set of predefined semantic concepts, the system applies content-based indexing techniques to generate probability scores for each image in a lifelog to contain an instance of each semantic concept. This process, detailed in section III-A, is performed off-line, whenever a new lifelog image set is added to the archive, or a new visual concept detector is implemented. The goal of the visual indexing, in the off-line phase, is to assign a probability score to each image $\times$ concept.

At query time, users provides the system a text query such as given with the LSAT task, using natural language. The query is mapped to the predefined visual concepts, and a set of concepts are selected based on their similarity to define the query terms, the scores of the selected concepts are merged to define the score of the image to the query. Temporal indexing is straightforward based on the temporal information comes from the device with each image. The temporal indexing will filter the images and remain those meets the query temporally. Finally, a the remained images are ranked according to their visual scores and returned as a result to the user query.

Our proposed framework has been design to be as generic as possible. It operates on any lifelog collection for which temporal information are available. It can also operate independently of the available set of visual concept models.

In this paper, we apply a manual mapping between query to the visual concepts, however our intention is to automate this mapping using NLP techniques.

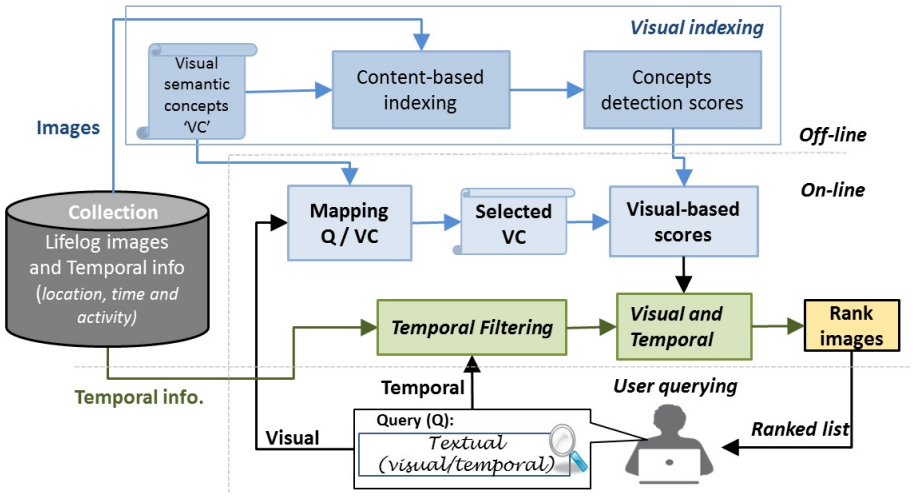


Fig. 1. Proposed lifelog semantic annotation framework

A. Visual indexing

The visual concept indexing used for the lifelog images is presented in Figure 2. It is 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 [5], namely the AlexNet network [6], the VGG network [13] and the GoogLeNet network [14]. Each of these networks have been learned on the ImageNet corpus. In order 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 [11], mainly we used the accelerated version of the MSVM [12], 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. Based on our previous experiments, on TRECVid [9] and Pascal-VOC [2] tasks, by merging the three descriptors the indexing system has a higher performance. Thus, the merging was applied as follow: the three vectors were optimized separately using the power-law and PCA approach [10], 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 TRECVid concept, 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 [12] 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).

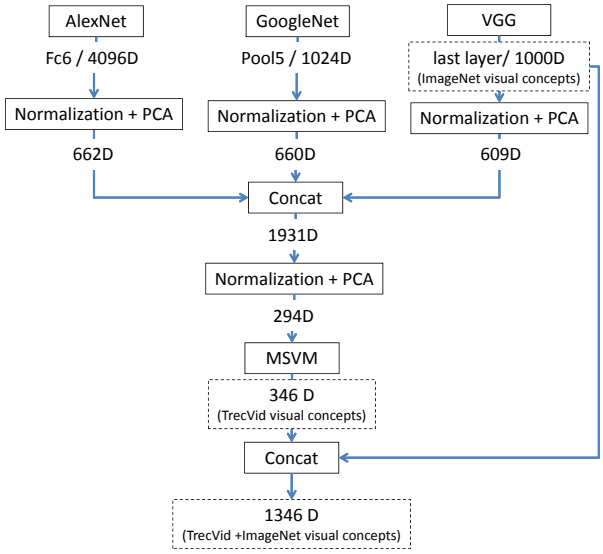


Fig. 2. Visual concept indexing

Given a topic "query", we link the terms of the topic manually to the set of ImageNet and TRECVid concepts. Though it would indeed be preferable to map fully automatically the

[2]http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=1

query on the available visual concepts, this is not really a problem in practice since the user building the queries could easily with no or very little extra cost directly build their queries using them. For the visual representation of the query, 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.

B. 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 I (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 I (bottom). These temporal concepts are binary, and describe each lifelog image.

TABLE I
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

C. Log indexing

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

D. Query processing

The query processing is currently 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 only afterward on the other information (temporal, activity, etc.). Theoretically however, it is like ANDing both aspects.

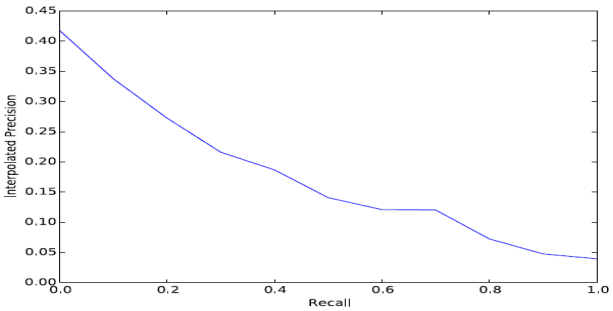


Fig. 3. Precision-Recall curve.

IV. OFFICIAL RESULTS DISCUSSION

A. 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). All the submitted runs 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 results according to the two levels and the two evaluation measures are presented in table II and the precision and recall curve is shown in fig. 3. Results are presented for our official submission including the use of 1000 concepts from VGG, of the 346 concepts from TRECVid and of all available metadata (time and location), as well as for three contrast conditions in which TRECVid concepts are removed (but metadata are kept), in which metadata are removed (all visual concepts are kept) and in which only metadata are used. The results obtained are very good, for a first participation in a Lifelog retrieval campaign, especially for the event-based mAP and nDCG evaluations. The contrast conditions show that TRECVid concepts bring a significant improvement over using only the 1000 ImageNet ones, that metadata alone do not give very good results but they give a significant boost when used in conjunction with visual concepts.

TABLE II
MRIM RESULTS OFFICIAL EVALUATION

metric	nDCG		mAP	
level	Event	Image	Event	Image
Official submission	0.3896	0.2455	0.2940	0.1667
Without TRECVid	0.2841	0.1808	0.2069	0.1249
Without metadata	0.2955	0.1509	0.1948	0.0916
Metadata only	0.0753	0.0783	0.0490	0.0561

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.

B. Details

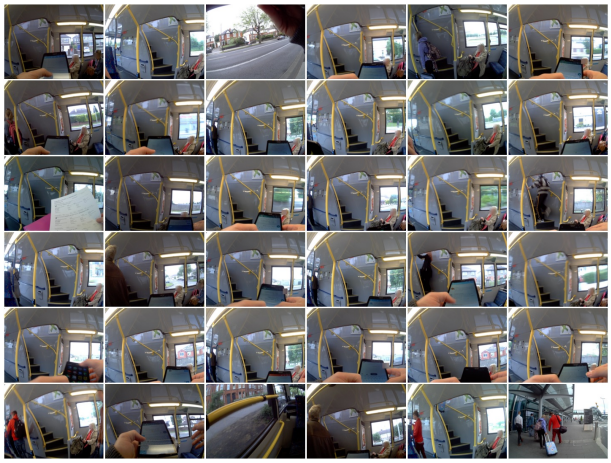
Among queries generated from the 48 initial topics provided by the task organizers, two of them, corresponding to topics 009 (i.e. *New Key*) and 048 (i.e. *Checkout*), led to an empty query. From the remaining 46 topics, we get the following statistics:

- 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 (i.e. *Lion at the Gate*) 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 (i.e. *A Movie on the Flight*) 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). 13 additional 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).

To discuss the visual indexing, figure 4 shows the top 30 images retrieved for two queries namely: *The Church* topic '014' (in top) and *Bus to the Airport* topic '031' (in bottom). As we can see that the retrieved images are very relevant visually to the searched topics, in a the images are taken inside a) church and in b) images are inside a bus with the temporal aspect we filtered the bus images to select only the images which are taken before arriving to an airport.

We discuss now the results query by query, by focusing on the event-based Average Precision results as presented in Figure 5. 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 5 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 them use

(a) Topic '014' (The Church).



(b) Topic '031' (Bus to the Airport).

Fig. 4. Screen shots of the top 30 retrieved images for topic '014' and '031'.

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 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 our overall results.

V. Conclusion

We proposed a way to retrieve events in a lifelog data stream. It relies on the use of both content-based and metadata-based information, all turned into concepts onto which query

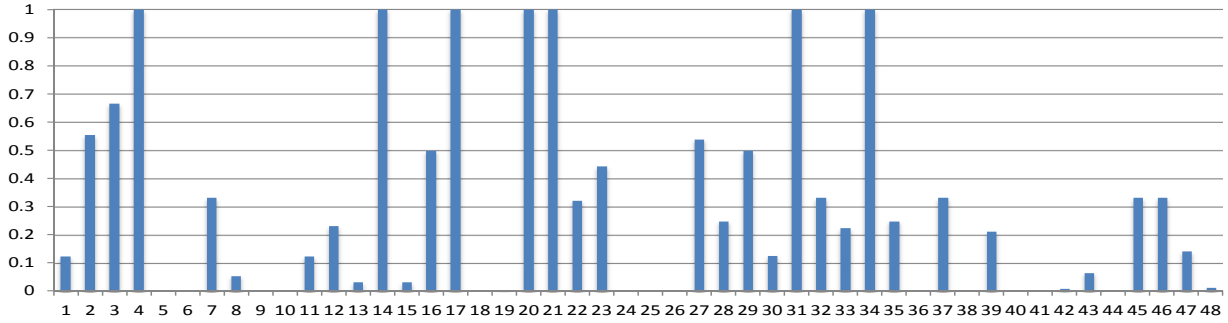


Fig. 5. Event-based AP official result per topic.

terms are mapped. For the content-based concepts, we used the ImageNet and TRECVid ones with detectors based on deep features. For the temporal one, we mapped time stamps to symbolic periods (e.g. "morning"). The location (e.g. home) and activity (e.g. walking) ones were directly extracted from the metadata where they were recorded by the sensors (with integrated processing).

Evaluation was conducted in the context of the NTCIR-12 Lifelog task (LSAT) where we obtained relatively good performances, we were ranked first at the official results, especially considering that we worked with data "from the wild". This performance comes mostly from the use of visual concepts that were quite accurately detected using deep learning-based techniques. 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 automatic processes (or direct query formulation based on the available concepts) will be able to achieve similar (or even better) results.

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) must also be studied 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. Part of the computations presented in this paper were performed using the Froggy platform of the CIMENT infrastructure (https://ciment.ujf-grenoble.fr), which is supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA) and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the programme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche.

References

[1] A. R. Doherty, N. Caprani, C. . Conaire, V. Kalnikaite, C. Gurrin, A. F. Smeaton, and N. E. OConnor. Passively recognising human activities through lifelogging. *Computers in Human Behavior*, 27(5):1948 – 1958, 2011. 2009 Fifth International Conference on Intelligent ComputingICIC 20092009 Fifth International Conference on Intelligent Computing.

[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. Ntcir lifelog: The first test collection for lifelog research. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy*. ACM, July 2016 (to be published).

[4] C. Gurrin, A. F. Smeaton, and A. R. Doherty. Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval*, 8(1):1–125, 2014.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] B. Safadi, N. Derbas, A. Hamadi, M. Budnik, P. Mulhem, and G. Quénot. LIG at TRECVid 2015: Semantic Indexing. In *Proceedings of TRECVid*, Gaithersburg, MD, United States, Nov. 2015.

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

[11] B. Safadi and G. Quénot. Evaluations of multi-learners approaches for concepts indexing in video documents. In *RIAO*, pages 88–91, 2010.

[12] 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.

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

[14] 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.

[15] L. Zhou, H. Joho, F. Hopfgartner, R. Albatal, and C. Gurrin. Ntcir12-lifelog, a test collection to support collaborative benchmarking of lifelog retrieval systems.