ImageCLEF 2017 LifeLog task

Duc-Tien Dang-Nguyen
Dublin City University
Dublin, Ireland
duc-tien.dang-nguyen@dcu.ie

Luca Piras
University of Cagliari
Cagliari, Italy
luca.piras@diee.unica.it

Michael Riegler
Simula Research Lab.
Oslo, Norway
michael@simula.no

Cathal Gurrin
Dublin City University
Dublin, Ireland
cgurrin@computing.dcu.ie

Giulia Boato
University of Trento
Trento, Italy
giulia.boato@unitn.it

Pål Halvorsen
Simula Research Lab.
Oslo, Norway
paalh@simula.no

Abstract

In this paper we present the ImageCLEF 2017 LifeLog task[1], which ran at ImageCLEF 2017[2]. We outline the dataset employed, along with the tasks, a task overview and the evaluation metodology of this task.

Introduction
The availability of a large variety of personal devices, such as smartphones, video cameras as well as wearable devices that allow capturing pictures, videos, and audio clips in every moment of our life is creating vast archives of personal data where the totality of an individual's experiences, captured multi-modally through digital sensors are stored permanently as a personal multimedia archive. These unified digital records, commonly referred to as *lifelogs*, gathered increasing attention in recent years within the research community. This happened due to the need for systems that can automatically analyse these huge amounts of data in order to categorize, summarize and also query them to retrieve the information that the user may need.
Despite the increasing number of successful related workshops and panels (e.g. JCDL 2015, iConf 2016, ACM MM 2016) lifelogging has seldom been the subject of a rigorous comparative benchmarking exercise as, for example, the new lifelog evaluation task at NTCIR-12[3]. This task aims to bring the attention of lifelogging to an, as wide as possible, audience and to promote research into some of the key challenges of the coming years.

Task Overview
The ImageCLEF 2017 LifeLog task aims to be a comparative evaluation of information access and retrieval systems operating over personal lifelog data. The task consists of two sub-tasks, both (or either) can be participated in independently. These sub-tasks are:
- Lifelog Retrieval Task (LRT)
- Lifelog Summarization Task (LST)

Lifelog retrieval task
The participants should analyse the lifelog data and according to several specific queries return the correct answers. For example:

- Shopping for Wine: Find the moment(s) when I was shopping for wine in the supermarket.
- Shopping for Fish: Find the moment(s) when I was shopping for fish in the supermarket.
- The Metro: Find the moment(s) when I was riding a metro.

[1] http://www.imageclef.org/2017/lifelog
[2] http://www.imageclef.org/2017
[3] http://ntcir-lifelog.computing.dcu.ie/NTCIR12/

Challenges in Machine Learning: Gaming and Education (CiML - NIPS 2016 workshop), Barcelona, Spain.

Number of Lifeloggers	3
Size of the Collection (Images)	88,124 images
Size of the Collection (Locations)	130 locations
Number of LRT Topics	16
Number of LsT Topics	5

Table 1: Statistics of Lifelog Dataset

The ground truth for this sub-task has been created by extending the queries from the NTCIR-12 dataset, which already provides a sufficient ground truth.

Lifelog summarization task
In this sub-task the participants should analyse all the images and summarize them according to specific requirements. For example:

> Public Transport: Summarize the use of public transport by a user. Taking any form of public transport is considered relevant, such as bus, taxi, train, airplane and boat. The summary should contain all different day-times, means of transport and locations, etc.

Particular attention should be paid to the diversification of the selected images with respect to the target scenario. The ground truth for this sub-task has been created utilizing crowdsourcing and manual annotations.

Dataset
The Lifelog dataset consists of data from three lifeloggers for a period of about one month each. The data contains a large collection of wearable camera images (circa two images per minute), an XML description of the semantic locations (e.g. Starbucks cafe, McDonalds restaurant, home, work) and the physical activities (e.g. walking, transport, cycling), of the lifeloggers at a granularity of one minute. A summary of the data collection is shown in Table 1.
Given the fact that lifelog data is typically visual in nature and in order to reduce the barriers-to-participation, the output of the CAFFE CNN-based visual concept detector is included in the test collection as additional metadata. This classifier provided labels and probabilities of occurrence for 1,000 objects in every image. The accuracy of the CAFFE visual concept detector is very variable, and is representative of the current generation of off-the-shelf visual analytics tools.

Topics
Aside from the data, the test collection includes a set of topics (queries) that are representative of the real-world information needs of lifeloggers. There are 16 and 5 ad-hoc search topics representing the challenge of retrieval for the LRT task and the challenge of summarization for the LST task, respectively.

Evaluation Methodology
For the *Lifelog Rerieval Task* evaluation metrics based on NDCG (Normalized Discounted Cumulative Gain) at different depths are used, i.e., NDCG@N, where N will vary based on the type of the topics, for the recall oriented topics N will be larger (>20), and for the precision oriented topics N will be smaller N (5 or 10 or 20).
In the *Lifelog Summarization Task* classic metrics will be deployed. These metrics are:
- Cluster Recall at X (CR@X) — a metric that assesses how many different clusters from the ground truth are represented among the top X results;
- Precision at X (P@X) — measures the number of relevant photos among the top X results;
- F1-measure at X (F1@X) — the harmonic mean of the previous two.

Various cut off points are to be considered, e.g., X=5, 10, 20, 30, 40, 50. Official ranking metrics this year will be the F1-measure@20 or images, which gives equal importance to diversity (via CR@20) and relevance (via P@20).
Participants are allowed to undertake the sub-tasks in an interactive or automatic manner. For interactive submissions, a maximum of five minutes of search time is allowed per topic. In particular, the organizers would like to emphasize methods that allow interaction with real users (via Relevance Feedback (RF), for example), i.e., beside of the best performance, the way of interaction (like number of iterations using RF), or innovation level of the method (for example, new way to interact with real users) are evaluated.