NTCIR Lifelog: The First Test Collection for Lifelog Research

Cathal Gurrin
Dublin City University, Ireland
cathal.gurrin@dcu.ie

Hideo Joho
University of Tsukuba, Japan
hideo@slis.tsukuba.ac.jp

Frank Hopfgartner
University of Glasgow, UK
frank.hopfgartner@glasgow.ac.uk

Liting Zhou
Dublin City University, Ireland
becky.zhou@dcu.ie

Rami Albatal
Heystaks Ltd., Ireland
rami.albatal@heystaks.com

ABSTRACT

Test collections have a long history of supporting repeatable and comparable evaluation in Information Retrieval (IR). However, thus far, no shared test collection exists for IR systems that are designed to index and retrieve multimodal lifelog data. In this paper we introduce the first test collection for personal lifelog data. The requirements for such a test collection are motivated, the process of creating the test collection is described, along with an overview of the test collection and finally suggestions are given for possible applications of the test collection, which has been employed for the NTCIR12-Lifelog task.

CCS Concepts

•**Information systems** → **Test collections;**

Keywords

lifelog, test collection, information retrieval, multimodal, evaluation

1. INTRODUCTION

One aspect of IR that has been gathering increasing attention in recent years is the concept of lifelogging. Lifelogging is defined as a form of pervasive computing, consisting of a unified digital record of the totality of an individual's experiences, captured multi-modally through digital sensors and stored permanently as a personal multimedia archive [3]. Lifelogging typically generates multimedia archives of life-experience data in an enormous (potentially multi-decade) lifelog. Therefore a lifelog needs to be organised and searchable to be of value to the individual / lifelogger, hence there have been calls for a test collection of lifelog data [6].

However, the design and construction of a lifelog test collection is not trivial. Jones and Teeven [9], in the context of personal information management (PIM), state that *"the design of shared test collections for PIM evaluation requires some creative thinking, because such collections must differ from more traditional shared test collections"*. Although an increased level of individual information has been incorporated into the design of recent test collections (e.g., Contextual Suggestion Track at TREC [2]), the scope of individual information represented by "user profiles" is far too limited to evaluate lifelog systems.

The contribution of this paper is a description of the requirements for a lifelog test collection (Section 2) and a description of the first lifelog test collection (Section 3) that has been employed in the Lifelog Task[1] of the NTCIR-12 Evaluation Forum. This is followed (Section 4) by a short list of potential applications of the test collection.

2. CREATING THE TEST COLLECTION

The basic structure of test collections used for IR research is based on the standard three components of 1) a collection of domain-representative documents, 2) a set of queries (called topics) that are representative of the domain of application, and 3) a set of relevance judgements. This section describes requirements that are particularly relevant to lifelog test collections. Moreover, we share our experiences of generating a test collection of personal lifelog data.

2.1 Requirements for the Test Collection

Prior to generating the test collection we defined requirements for the collection based on our experiences of developing prototype lifelogging applications and relevant literature concerning lifelogging and human memory, such as [11].

Validity of the Collection.

In order to allow for statistically significant studies in the field of lifelogging, a test collection needs to contain (a sample of) a dataset that is large enough to represent real-world data of lifeloggers. Since many lifeloggers use visual data to record daily activities and reflect upon their behaviour (e.g., [1]), we needed visual lifelogs captured from wearable cameras along with metadata describing the daily life of the lifelogger, such as locations and physical activities. Similarly, realistic topics are required representing real-world information needs, based on the experience of individuals who engage in lifelogging.

[1]http://ntcir-lifelog.computing.dcu.ie/ (Visited: 25 Mar 2016)

SIGIR '16, July 17-21, 2016, Pisa, Italy

ACM ISBN 978-1-4503-4069-4/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2911451.2914680

Number of Lifeloggers	3
Size of the Collection (GB)	18.18GB
Size of the Collection (Images)	88,124 images
Size of the Collection (Long-stay Semantic Locations)	130 locations
Size of the Collection (Visual Concept Metadata)	825MB
Size of the Collection (Visual Concepts Detected)	1,000
Number of Known Item (LSAT) Topics	48
Number of Insight Analytics (LIT) Topics	10

Table 1: Statistics of NTCIR-12 Lifelog Data

Privacy by Design.

In the field of lifelogging, personal sensor data (especially camera or audio data) will carry privacy concerns [10, 4]. Therefore, we must consider the principles of privacy-by-design when creating the test collection. These includes removing user-identifiable data from the collection, yet maintaining the usefulness of the data for the proposed purpose of supporting experimental IR research.

Facilitation of Research Activities.

In order to encourage comparative evaluation over lifelog data, the dataset needs to consist of sufficient metadata so as to lower the overhead for participation, hence allowing researchers interested in a broad range of IR-related applications to utilise the test collection. Moreover, a reusable test collection is required that can support a number of years of ongoing research activities. Addressing these issues, a lifelog dataset requires a set of relevance judgements that can be utilised both as a source of data for comparative evaluation as well as being later utilised as a source of training data for future lifelog system experimentation.

2.2 Process of Creating the Test Collection

Given the requirements for the lifelog test collection to be as realistic as possible, there were a number of steps that were taken to ensure that this was the case, while at the same time respecting the concept of privacy-by-design and the personal nature of the rich media data being donated by the lifeloggers.

Low-overhead Multimodal Data Capture.

Due to the long-term, always-on, nature of lifelog data gathering, it was important to reduce the overhead on the lifelogger of gathering the data. Hence, the data was gathered using only two logging devices, the OMG Autograper wearable camera and a smartphone running the Moves app[2]. The OMG Autographer is worn on a lanyard around the neck, is orientated towards the activity the wearer is engaged in and can operate for a full-day on a battery charge. This camera takes photos passively (i.e., without explicit user intervention) and as such, it gathers a detailed digital trace of the activities of the wearer at about two images per minute. This camera is a later generation of the Microsoft Sensecam wearable camera [7]. The Moves app is a smartphone app that automatically records user activity in terms of semantic locations and physical activities (e.g. waking, cycling, running, transport), without requiring any user intervention. This app was installed on the personal smartphones used by the lifeloggers. The moves data was exported from the Moves cloud-service after the data gathering process was complete.

[2]https://moves-app.com, Visited 11 May 2016

Temporal Alignment.

It was important to ensure temporal alignment of the sensor data, given that it is from two distinct devices. It was necessary to check and resolve alignment problems (typically in the order of 1-2 minutes) for one lifelogger by cross-referencing reported timestamps from the Autographer camera with clocks captured daily in the real-world.

Data Filtering for Privacy Preservation.

Given the personal nature of lifelog data, it was necessary to give the lifeloggers an opportunity to remove any data that they may be uncomfortable sharing. This involved a manual inspection of all their lifelog data before sharing it with us. Following this, all images were reviewed by one trusted individual with oversight of the entire collection to ensure that no potentially embarrassing or offensive images were accidentally included in the collection.

Anonymisation of the Dataset.

Two steps were taken to ensure privacy of both the lifeloggers and individuals (subjects and bystanders) captured in the lifelog, by removing identifiable content. Firstly, each recognisable face in every image was blurred in a manual process, which ensured no false positives or missed faces. A second step was to resize every image down to 1024 x 768 resolution which had the effect of both reducing the disk-size of the collection, but also rendering the majority of any on-screen text captured by the lifelogging camera to be illegible. Since privacy extends beyond faces in images, the Moves app automatically converts all locations from absolute GPS locations into semantic locations, which resulted in potentially sensitive absolute addresses being labeled with generic names such as 'home' or 'work', thereby making it more unlikely that the lifeloggers could be identified.

Figure 1: Sample Wearable Camera images from the Test Collection

3. DETAILS OF THE TEST COLLECTION

3.1 The Dataset

The NTCIR Lifelog test collection consists of data from three lifeloggers for a period of about one month each. The lifeloggers would all have had involvement with a university, but in different roles. The lifeloggers gathered data in an all-day gathering process; thereby gathering a wide range of daily activities. The data volume was roughly equally distributed per lifelogger. The data consists of a collection of wearable camera images (taken by an OMG Autographer camera) as shown in Figure 1; in total there were 88,124 images in the test collection. It also contains 130 semantic locations (e.g., Starbucks cafe, Dublin Airport, home, work) which were the places where the lifeloggers went (and lingered for some time) during their month of data gathering. Additionally, the physical activities (e.g., walking, transport, cycling) of the lifelogger during this month were included on a minute-by-minute basis. All data is accompanied by XML markup at the minute level of granularity. An example of the XML description is shown in Figure 2.

```
<minute id="906">
  <location>Work</location>
  <activity>walking</activity>
  <images>
   <image>
    <image-path>/u1/2015-02-18/
     20150218/150615e.jpg</image-path>
    <image-id>u1_2015-02-18_150615_1
    </image-id>
   </image>
   <image>
    <image-path>/u1/2015-02-18/
     20150218/150652e.jpg</image-path>
    <image-id>u1_2015-02-18\150652_2
    </image-id>
   </image>
  </images>
</minute>
```

Figure 2: An example of the XML data description for one minute

Given the fact that lifelog data is typically visual in nature, and to lower the barriers to participation for non-computer vision researchers, the output of the CAFFE visual concept detector [8] was included in the test collection as a source of additional visual metadata. This software provided labels and probabilities of occurrence for 1,000 objects in every image (objects such as car, pizza, desktop computer). A summary of the test collection is shown in Table 1.

3.2 The Topics

Aside from the data, the test collection includes a set of topic descriptions that are representative of the real-world information needs of lifeloggers and represent the *Retrieval* and *Reflection* reasons for accessing memories [11]. Some topics have a single correct result, whereas others require a number of relevant events to be returned for every topic. The 48 topics, called LSAT (Lifelog Semantic Access) Topics, were suggested by the three lifeloggers and they represent the challenge of *Retrieval* from memories. These topics were evaluated in terms of traditional IR effectiveness measurements such as Precision, Recall, MAP and NDCG. An example of an LSAT topic is shown in Figure 3. Table 2 shows ten example LSAT topics, the number of relevant results for each topic, and the recall performance of the best official automatic and interactive runs. For further details, the reader is referred to the Lifelog task overview paper from the NTCIR-12 proceedings [5].

TITLE: Tower Bridge
DESCRIPTION: Find the moment(s) when I was looking at Tower Bridge in London.
NARRATIVE: To be considered relevant, the full span of Tower Bridge must be visible. Moments of crossing the Tower Bridge or showing some subset of Tower Bridge are not considered relevant.

Figure 3: LSAT Topic Example

Topic Title	Relevant	Automatic	Inter
Photographing a Lake	2	2	0
Tower Bridge	1	1	1
Driving a Rental Car	19	0	5
Lost	1	1	N/A
Man in a Burberry Coat	1	1	0
Antiques Store	3	3	3
Building a Computer	14	1	1
Bus to the Airport	1	1	1
Checking the Menu	3	0	0
Car Repair	1	0	1
Playing Lottery	1	1	0

Table 2: Ten Sample Topics from the NTCIR-12 Lifelog Collection, showing the number of relevant events, the recall performance of the best performing Automatic and Interactive runs.

Additionally, there are ten exploratory topics representing the challenge of supporting *Reflection* from memories. These are called LIT (Lifelog Insight) Topics and are not evaluated in a traditional sense, rather, participants were encouraged to prepare insights (in any form) and demonstrate them directly to other participants at the NTCIR-12 Conference. An example of an LIT topic is shown in Figure 4.

TITLE: Early Morning Commute
DESCRIPTION: Early Provide insights on the methods of, and duration, each lifelogger spends commuting to work.
NARRATIVE: Commuting to work or university, via whatever means, is relevant. Commuting to a meeting in a location that is not the user's normal place of work is also relevant if it could be considered to be a morning commute to work. Commuting home is not relevant. General travelling is not relevant.

Figure 4: LIT Topic Example

3.3 The Relevance Assessments

The pooling method of creating relevance judgements is the typical approach for large-datasets in IR. Specifically with regard to the LSAT task, however, the relevance judgements were manual (non-pooled), and were generated by the lifeloggers and reviewed by the task organisers. The data was segmented into topic-specific events manually and at evaluation time, submissions in the form of image IDs (smallest unit of retrieval) were mapped onto events and the events judged as relevant or not relevant. If there were more than one image from any given event identified as relevant, then only the top ranked image (from that event) was selected for evaluation.

In addition, so as to make the test collection as useful as possible for a wide-range of researchers, we also provided relevance judgements for the LSAT task on a per-image basis, as the smallest unit of retrieval possible from this test collection.

Since the LIT task was an exploration-focused task, rather than a retrieval-focused task, there were no relevance judgements prepared for the LIT task. Instead, the LIT task participants presented their findings, techniques and insights in oral and poster session at NTCIR-12.

4. POTENTIAL RESEARCH CHALLENGES

Having a large collection of annotated personal data, such as this lifelog test collection, opens up a number of research opportunities:

- Multi-modal search and retrieval over archives of personal data.
- Lifelog-specific access, addressing many applications of memory, as defined by [11].
- Activity recognition from real-world data, in terms of both physical and real-world activities.
- Visual concept extraction, from real-world all-day wearable camera data (with additional supporting metadata).
- Time-series analysis of all-day personal data over extended time-periods from one, or multiple users.
- Generating insights & analytics from real-world datasets of wearable personal sensing.
- Contextual analysis of real-world user activities, to support exploratory approaches to contextual information access.
- Privacy-aware retrieval, to explore the privacy concerns surrounding search and retrieval from large lifelog archives.

The NTCIR-12 Lifelog test collection can be used to facilitate many of the above research challenges of lifelog data. Both the document collection and query data will be released to research communities after the NTCIR-12 conference and will be made available to researchers who sign up to a (standard) data-release agreement.

5. CONCLUSION

This paper reported the design and construction of the first test collection for lifelog research. A document collection and information needs of the NTCIR Lifelog test collection are highly individual and multimodal when compared to conventional test collections.

Research conducted by participants of NTCIR-12 Lifelog Task was presented at the NTCIR-12 Conference, June 7-10, 2016 at Tokyo, Japan, and at a parallel workshop at the University of Glasgow. Based on our experiences and feedback from participants, we will prepare a new test collection for NTCIR-13, which will incorporate additional sources of contextual data (e.g., audio, computer interactions, physical accelerations/movement, biometrics) and the provision of more raw-metadata, as opposed to the semantic data that was provided with this collection.

6. ACKNOWLEDGMENTS

We acknowledge the financial support of Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289 and the input of the DCU ethics committee and the risk & compliance officer.

7. REFERENCES

[1] S. Chowdhury, P. J. McParlane, M. S. Ferdous, and J. Jose. "my day in review": Visually summarising noisy lifelog data. In *ICMR'15*, pages 607–610. ACM, 2015.

[2] A. Dean-Hall, C. L. A. Clarke, J. Kamps, P. Thomas, and E. M. Voorhees. Overview of the TREC 2014 contextual suggestion track. In *TREC'14*, 2014.

[3] M. Dodge and R. Kitchin. "Outlines of a world coming into existence": Pervasive computing and the ethics of forgetting. *Environment and Planning B*, 34(3):431–445, 2007.

[4] C. Gurrin, R. Albatal, H. Joho, and K. Ishii. A privacy by design approach to lifelogging. In *Digital Enlightenment Yearbook 2014*, pages 49–73. IOS Press, 2014.

[5] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal. Overview of NTCIR-12 lifelog task. In *Proceedings of NTCIR-12 Conference*. NII, 2016.

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

[7] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. R. Wood. Sensecam: A retrospective memory aid. In *UbiComp 2006*, pages 177–193, 2006.

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, pages 675–678, 2014.

[9] W. Jones and J. Teevan. *Personal information management*. University of Washington Press, 2011.

[10] N. Li and F. Hopfgartner. To log or not to log? SWOT analysis of self-tracking. In *Lifelogging - Interd. Approaches to Unravel the Phenomenon of Digital Self-Tracking*. Springer VS, 2016.

[11] A. J. Sellen and S. Whittaker. Beyond total capture: A constructive critique of lifelogging. *Commun. ACM*, 53(5):70–77, May 2010.