# A Baseline Search Engine for Personal Life Archives

Liting Zhou
Insight Centre for Data Analytics
Dublin City University
Dublin, Ireland
zhou.liting2@mail.dcu.ie

Duc-Tien Dang-Nguyen
Insight Centre for Data Analytics
Dublin City University
Dublin, Ireland
duc-tien.dang-nguyen@dcu.ie

Cathal Gurrin
Insight Centre for Data Analytics
Dublin City University
Dublin, Ireland
cgurrin@computing.dcu.ie

## ABSTRACT

In lifelogging, as the volume of personal life archive data is ever increasing, we have to consider how to take advantage of a tool to extract or exploit valuable information from these personal life archives. In this work we motivate the need for, and present, a baseline search engine for personal life archives, which aims to make the personal life archive searchable, organizable and easy to be updated. We also present some preliminary results, which illustrate the feasibility of the baseline search engine as a tool for getting insights from personal life archives.

## CCS CONCEPTS

•**Information systems → Digital libraries and archives; Personalization;** *Users and interactive retrieval; Evaluation of retrieval results;*

## KEYWORDS

Search Engine; Lifelogging; Personal Life Archive

## 1 INTRODUCTION

Within the last decade, we have witnessed an increase in both the processing power and prevalence of wearable computing devices. With over 100 million wearable devices sold in 2016[1], it is clear that the current decade is the wearables decade, which follows on from the mobile/social network era in the previous decade. The combination of recent advances in wearable/mobile computing and information search technologies have each helped to bring us to the point where any individual with a cell-phone or a custom off-the-shelf device can now begin to gather vast volumes of personal information in their own personal life archives or also called lifelogs, which may currently be spread across multiple devices and services. These personal life archives can contain information about every activity an individual participates in, such as where they go, how they get there, who they speak with, what they see and what information they access, in short, a complete onmi-present digital life diary is now possible to generate for every individual who chooses to do so. This is made possible by the recent development of power-efficient sensing devices, an enabling technology which is embedded in the modern cell-phone or any other devices.

Along with the advantages, lifeloggers are typically facing the problem of overwhelming amounts of data [14]; essentially a modern form of the art-museum problem of the initial hypermedia systems. Consequently, there is a need for technologies that can transform the non-indexed personal life archive into a form that is organised and searchable. However, there are a number of requirements for such life archives. The first one is concerned with privacy and ethical considerations concerns [4, 13]. This issue has implications for both the individual and society since the personal life archive contains a digital trace of everything of one's experiences, including identifiable individuals. Assuming such problems can be addressed, then it becomes necessary to organise the archive so that it can be a useful knowledge-source for the individual. This can be achieved by i) segmenting it into a series of distinct events or activities [7], ii) automatically labeling those events from both the content [6] and context [11], iii) automatically detecting interesting content (e.g. people) and event novelty to identify those events that are more interesting to reflect on [8], iv) presenting segments of this personal life archive to the user as required and v) analysing this data to provide new knowledge to the user. These five issues can be addressed once there is an efficient technology that allows an individual to retrieve the basic moments from the personal life archives in a reliable and accurate manner; this is the second challenge of personal life archives. In order to overcome those challenges, in this paper, we propose a baseline search engine for personal life archives, which aims at addressing these two requirements by providing the technology that can index and organise the personal life archive, while keep the raw data private and protected.

The main idea behind this baseline search engine is first to provide a starting point for researchers in the area, as well as a documented system for comparative analysis. The baseline search engine must extract features from the raw data for higher level analysis. Then, it indexes all the extracted features and hierarchically organises them which allow queries to be processed via a simple API/interface. Details of this system will be introduced in the next section. Later, some applications will be presented in Sections 3 and 4 before the conclusion in Section 5.

## 2 THE PROPOSED SYSTEM

As with any information retrieval system, search and ranking is only one aspect of the challenge. When we are developing retrieval systems, whether for personal life archives or for WWW pages, to find a starting point, we need to understand how the personal life archive will be accessed. Let us begin by considering how people access their own digital photo or video archives. For sufficiently

[1] goo.gl/Z76xk3


*LTA'17, October 23, 2017, Mountain View, CA, USA*

ACM ISBN 978-1-4503-5503-2/17/10…$15.00
DOI: https://doi.org/10.1145/3133202.3133206

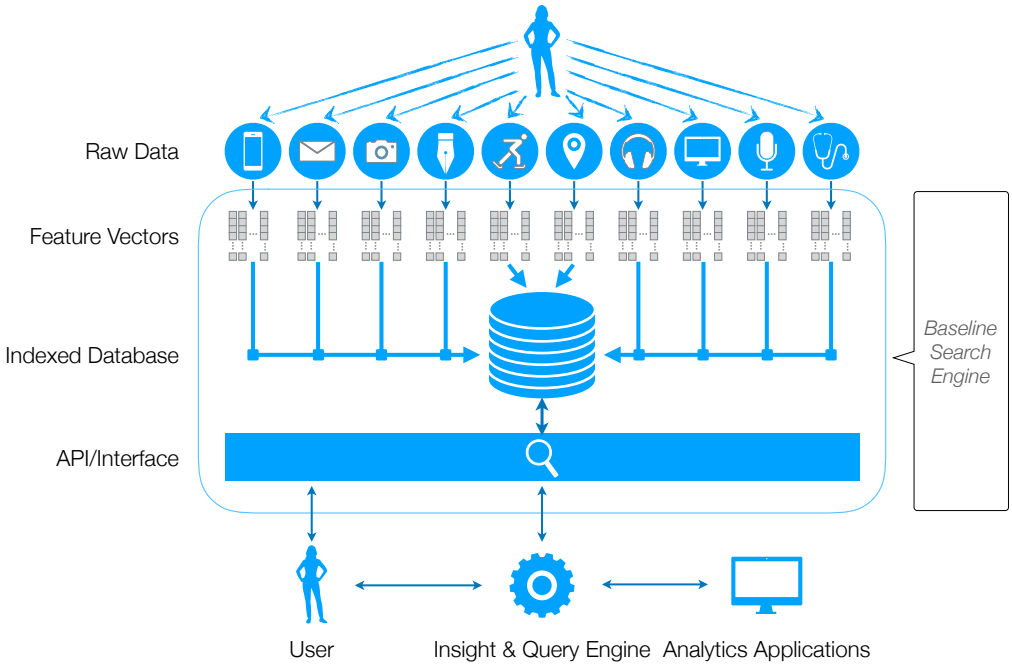


**Figure 1: Baseline search engine.**

small archives, a browsing mechanism (manually or automatic organization into clusters based on folders or events) is acceptable, where the selection of an axis of browsing results in the generation of a manageable set of result documents. Consider browsing a photo archive using date or location. However, when the archives become larger and less organised, a search or search/browse metaphor is normally chosen to support fast and effective access. This is where we position the baseline search engine, as a basic building-block to support the searching of life archives.

Figure 1 summarises how the proposed baseline search engine system works, as follows: Personal life archives can be gathered from the phone data, emails, cameras, written notes, activities, locations, music, computer usage, speech, biometric information and many other sources. This raw data is firstly transformed into feature vectors, which are measurable, searchable, and more importantly: allows the raw data to be private and protected. These feature vectors are then indexed and hierarchically organized. Finally, users or other insight and query engines can retrieve information via a simple API/Interface which simply returns the the indexed data that matched with the queried criteria.

## 2.1 Feature Extraction

There are many different kinds of information in a personal life archive, and each of them requires different techniques for feature extraction. For example, if the information is a text-based document, it can be transformed and indexed into keywords and their associated term / document frequencies. If the information is the image taken by the wearable camera, it can be the visual features or the concepts detected from the content of the image. Shown in Figure 2 is an example of how a picture taken by a wearable camera is turned into features which are the colors, the tags, and the predicted caption using a visual annotation tool.

## 2.2 Data Organisation

The human memory system has evolved over millions of years to store autobiographical memories, and we believe that personal digital archives needs to engage some form of biomimicry that models the digital archive search on the human memory mechanism. Past literature has motivated that the human mind stores information in distinct events or episodes, that similar episodes are associated with each other, and that more important episodes are more strongly remembered [8]. However in addition to this, given the range of data streams associated with a given important episode of interest, a short descriptive summary narrative of that episode is required. As a starting point, the personal digital archive information could be arranged as follows:

*Raw data should be hierarchically arranged and stored:* Typically in information retrieval (IR), there is a single basic unit for indexing and retrieval, typically called the document. For many IR tasks, this basic unit is the preferred as unit of retrieval and choosing the basic unit is usually trivial. With lifelog data, it is not trivial to decide what the basic unit is since the lifelogs are multimodal with different modes captured at different frequencies (1 second to potentially 1 day) [4]. In order to deal with this problem, we followed the study in [4] that sorts the data by chronological order, and use the minute as the basic units (as shown in Figure 3).

Building up from this basic unit, we organize the data at higher level which can be turned into more useful information. Typically, in a full day, we know that a person encounters anything upwards of 20 individual *episodes or events*, with each lasting about 30 minutes,

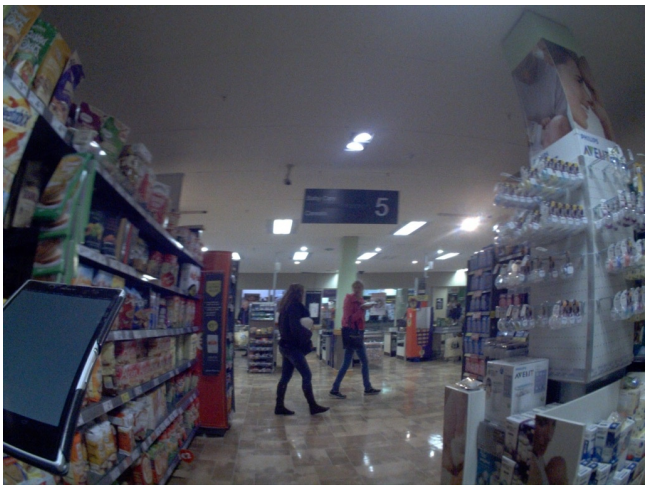

**Figure 2: An example of the results from Microsoft Vision API: "color": {"dominantColorForeground": "Grey", "dominantColorBackground": "Grey", "accentColor": "4B6980", "isBWImg": false, }; "tags": "indoor", "standing", "man", "room", "people", "woman", "table", "refrigerator", "kitchen", "holding", "young", "group", "playing", "large", "living", "video", "display", "store", "game"; "captions": { "text": "a group of people standing in a store", "confidence": 0.616320133 };**

though there is a lot of variety [5, 12]. Thus, we propose to organize the data hierarchically where the basic unit is the minute, but also including the episode or event nodes at a higher level, and ultimately leading to larger units such as days or multi-day events (e.g. holidays).

## 3 PRELIMINARY RESULTS

As a preliminary step, we apply our preliminary version [15] of the proposed baseline search engine to solve the problems in ImageCLEFlifelog 2017 [3] of ImageCLEF 2017 [10]. ImageCLEFlifelog 2017 gives some basic queries on lifelogs retrieval and asks the participant to return the relevant moments. An example of a query is:

"Find the moment(s) in which user u1 was using his laptop outside the work place".

In [15], we exploit the preliminary baseline search engine to solve the problem by "translating" the query into specific required pieces of information, i.e., smaller inquiries, and sent to the baseline search engine. For example, the sub-inquiries "translated" from the mentioned queried are:

User = {u1}, Concepts = {laptop}, Activity = {working, watching}, and Location = Anywhere\{'work'}

How to convert a topic into precise criteria for retrieval is the key question. It can be automatically done by considering any word in the topic as the queried concepts and then searching for all segments that contain those concepts; by applying natural language processing techniques; or by fine-tunning by a human in the loop, i.e., the user will read the topic and "translate" it into the search criteria. For ImageCLEFlifelog 2017, we tried with these all three solutions and obtained the best results of 39% accuracy with the fine-tunning method, measured by the normalized discounted cumulative gain (NDCG) at the top-ten, the official metric of the task. This result is considered to be very good performance for this kind of task.

We also built and made this proposed baseline search engine open access at: http://search-lifelog.computing.dcu.ie/, which are now serving as the basic tool for NTCIR-13 - Lifelog 2 task[2].

## 4 POTENTIAL APPLICATIONS

The potential for personal life archives is enormous. We do acknowledge that there are challenges to be overcome, such as data storage, security of data, and the development of a new generation of search and organisation tools based on the proposed baseline search engine, but we believe that these will be overcome and that we are on the cusp of a positive turning point for society; the era of the quantified individual who knows more about the self than ever before, has more knowledge to improve the quality of their own life and can share life events and experiences in rich detail with friends and contacts.

We firmly believe that personal life archive search engine will very soon be a phenomenon available to everyone and exploiting the personal life archives will positively impact on everyone who uses the technology. Through integration with other wearable sensors via for example Bluetooth on cell phones, this baseline search engine brings some potential applications and research problems, as follow:

- How can we identify and extract important information? Deciding what should be extracted from the data that is important for the users is a nontrivial task. Going beyond standard analysis like detecting visual objects or text will be important, forcing researchers to think creatively and go beyond simple analysis. Many research questions arise, such as, how to combine information from sensors with videos or images, or how to efficiently process the data. Also an important part here is to explore how context can be taken into account to improve the quality of the analysis.
- How can information and data be processed efficiently? Systems that have to process a huge amount of data in a complex way have to be efficient to make them useful to the users. This comes with challenges for multimedia systems researchers in terms of how to parallelize and process data efficiently in a reasonable amount of time.
- Most memory research exploiting visual lifelogging has focused on retrospective episodic memory tasks [2]. However as real-time upload becomes feasible, we can now begin to focus on supporting the human prospective memory system [1] by including rich user context. Given a user's prior set of experiences and preferences (historical lifelog data), and their current contextual situation (gathered via on-board mobile device and wearable sensors), prospective memory prompts can be provided by building ontop of this baseline system.
- The era of personal life archives will provide information to the individual about their own life activities; information that otherwise would go unnoticed. One can identify trends and pattern in lifestyle and wellness over an extended period of time

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

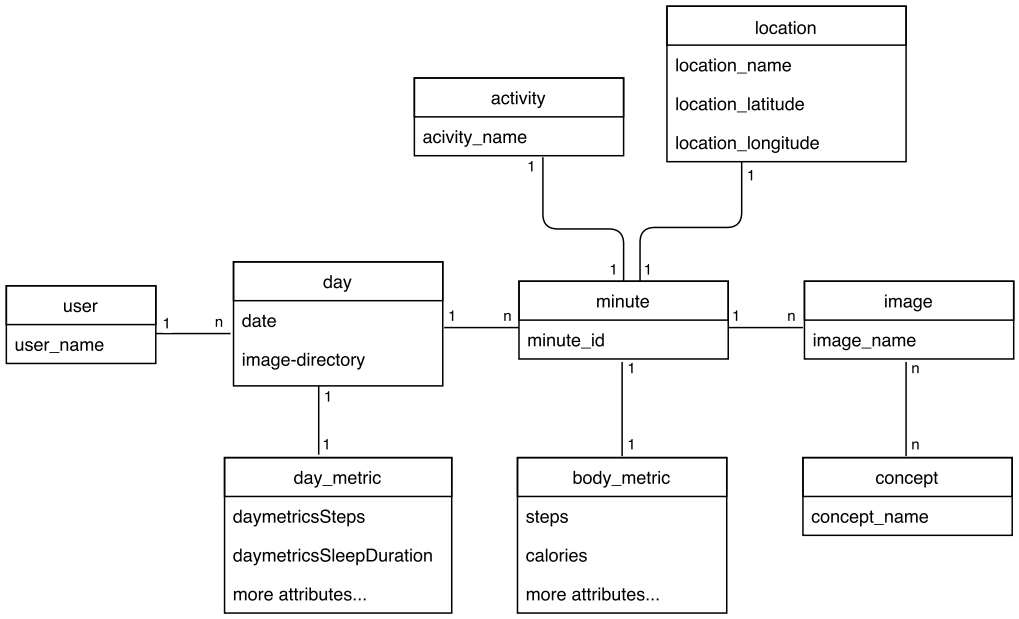


**Figure 3: An example of how the personal life archive data is organized.**

- Although we have spoken about the personal nature of life archive, there is likely to be enormous potential for mankind when the life archives of whole populations can be analysed together for trends and statistics based on the proposed search engine. To take just one example, in the spirit of the Framington study [9], analysis by reputable and trusted organizations of anonymised archives of populations could potentially lead to greater understanding of potential root causes of diseases and illnesses, on a scale heretofore unimaginable.

## 5 CONCLUSIONS

In this paper, we introduced a baseline search engine, a system that allows to retrieve the basic moments from the personal life archives in a reliable and efficient manner, while keeping the personal life archive private and protected. This baseline search engine can potentially be used to help user fully understand themselves to improve their life. In the future work, this baseline search engine will be enriched, and extended with more agile and advanced solutions, which aims to give better information for higher level of insight and query engines.

## 6 ACKNOWLEDGE

This publication has emanated from research supported in party by research grants from Irish Research Council (IRC) under Grant Number GOIPG/2016/741.

## REFERENCES

[1] Alan Baddeley (Ed.). 2004. *Your Memory: A User's Guide.* Carlton Books.
[2] Martin A. Conway and Catherine Loveday. 2011. SenseCam: The Future of Everyday Memory Research. *Memory* 19, 7 (Oct 2011), 685–807.
[3] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Giulia Boato, Liting Zhou, and Cathal Gurrin. 2017. Overview of ImageCLEFlifelog 2017: Lifelog Retrieval and Summarization. In *CLEF2017 Working Notes (CEUR Workshop Proceedings).* CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland.
[4] Duc-Tien Dang-Nguyen, Liting Zhou, Rashmi Gupta, Michael Riegler, and Cathal Gurrin. 2017. Building a Disclosed Lifelog Dataset: Challenges, Principles and Processes. In *Content-Based Multimedia Indexing (CBMI).*
[5] A.R. Doherty and A.F. Smeaton. 2008. Automatically segmenting lifelog data into events. In *Ninth International Workshop on Image Analysis for Multimedia Interactive Services.* IEEE, 20–23.
[6] Aiden Roger Doherty, Niamh Caprani, Ciaran O Conaire, Vaiva Kalnikaite, Cathal Gurrin, Alan F. Smeaton, and Noel E. O Connor. 2011. Passively recognising human activities through lifelogging. *Computers in Human Behavior* 27, 5 (2011), 1948–1958.
[7] Aiden R. Doherty, Chris J.A. Moulin, and Alan F. Smeaton. 2011. Automatically assisting human memory: A SenseCam browser. *Memory* 7, 19 (2011), 785–795.
[8] A R Doherty, K Pauly-Takacs, N Caprani, C Gurrin, C J A Moulin, N E OConnor, and A F Smeaton. 2012. Experiences of Aiding Autobiographical Memory Using the SenseCam. *Human-Computer Interaction* 27, 1-2 (2012), 151–174.
[9] Tavia Gordon, William P. Castelli, Marthana C. Hjortland, William B. Kannel, and Thomas R. Dawber. 1977. Predicting Coronary Heart Disease in Middle-Aged and Older Persons. *JAMA: The Journal of the American Medical Association* 238, 6 (1977), 497–499.
[10] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Helbert Arenas, Giulia Boato, Duc-Tien Dang-Nguyen, Yashin Dicente Cid, Carsten Eickhoff, Alba Garcia Seco de Herrera, Cathal Gurrin, Kovalev Vassili Islam, Bayzidul and, Vitali Liauchuk, Josiane Mothe, Luca Piras, Michael Riegler, and Immanuel Schwall. 2017. Overview of ImageCLEF 2017: Information extraction from images. In *CLEF 2017 Proceedings.* Springer, Dublin, Ireland.
[11] Ramesh Jain and Pinaki Sinha. 2010. Content without context is meaningless. In *Proceedings of the international conference on Multimedia (MM '10).* ACM, New York, NY, USA, 1259–1268.
[12] Daniel Kahneman, Alan B. Krueger, David A. Schkade, Norbert Schwarz, and Arthur A. Stone. 2004. A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science* 306 (2004), 1776–1780.
[13] David H. Nguyen, Gabriela Marcu, Gillian R. Hayes, Khai N. Truong, James Scott, Marc Langheinrich, and Christof Roduner. 2009. Encountering SenseCam: personal recording technologies in everyday life. In *Ubicomp '09: Proceedings of the 11th international conference on Ubiquitous computing.* 165–174.
[14] Abigail J. Sellen and Steve Whittaker. 2010. Beyond total capture: a constructive critique of lifelogging. *Comm. ACM* 53, 5 (2010), 70–77.
[15] Liting Zhou, Luca Piras, Michael Rieger, Giulia Boato, Duc-Tien Dang-Nguyen, and Cathal Gurrin. 2017. Organizer Team at ImageCLEFlifelog 2017: Baseline Approaches for Lifelog Retrieval and Summarization. *CLEF working notes, CEUR* (September 11-14 2017).