

# ENHANCING THE EFFECTIVENESS OF INFORMATION RETRIEVAL TECHNIQUES FOR KNOWN-ITEM RETRIEVAL FROM LIFELOGS

Liting Zhou, B.Sc.

A Dissertation submitted in fulfilment of the  
requirements for the award of  
Doctor of Philosophy (Ph.D.)

to the



Dublin City University

Faculty of Engineering and Computing, School of Computing

Supervisors

*Assoc. Prof.* Cathal Gurrin


*Prof.* Alan Smeaton

*Prof.* Chris Moulin (Université Grenoble Alpes)

February 2021

# Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Sign:   
(*Liting Zhou*)

Student No.: 16213671

Date: 25/02/2021



# Acknowledgements

I express my heartfelt gratitude to my internal supervisors Assoc. Prof. Cathal Gurrin, Prof. Alan Smeaton and external supervisors Prof. Chris Moulin (Université Grenoble Alpes), Assoc. Prof. Duc Tien Dang Nguyen (University of Bergen) for their valuable guidance not only on research topics but also on research methodologies. Their profound knowledge and innovative thoughts about lifelogging is nothing but impressive and enlightening for my PhD work.

Thanks to the main financial support from Irish Research Council (IRC) for the research in this thesis. The work of this PhD research has emanated from research conducted with the financial support of IRC under grant number "GOIPG/2016/741".

I would like to thank my external and internal examiners for their valuable feedback, Dr. Monica Landoni and Pr. Mark Roantree. Thanks to Dr. Jiang Zhou and Dr. Joseph Antony, who taught me Deep Learning Knowledge and also gradually introduced me to the topics which I became engaged with later. Thanks to my dear group members, who helped me build the annotation platform and complete the annotation work.

There are no words that I can say to express my love and appreciation to my parents for bringing me up and giving me the best education that I could have by being paradigmatic parents. Also, I can't forget to mention my great siblings who supported me through the toughest days.

I would like to extend my heartfelt thanks to everyone who helped and supported me throughout my PhD. Thank you everyone for steering me in the right direction.

# Table of Contents

<b>Table of Contents</b>	<b>i</b>
<b>List of Figures</b>	<b>iv</b>
<b>List of Tables</b>	<b>vi</b>
<b>List of Abbreviations</b>	<b>vii</b>
<b>Abstract</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Lifelog: Digital Diary . . . . .	1
1.1.1 Related Terminology . . . . .	2
1.1.2 Lifelogging Challenge . . . . .	3
1.2 Research Motivation and Challenges . . . . .	8
1.2.1 Capture & Storage Challenge . . . . .	9
1.2.2 Organisation Challenge . . . . .	11
1.2.3 Indexing and Annotation Challenge . . . . .	12
1.3 Hypothesis and Research Questions . . . . .	14
1.4 Research Contributions . . . . .	17
1.5 Thesis Outline . . . . .	17
<b>2 Related Work and Background</b>	<b>19</b>
2.1 Introduction . . . . .	19
2.2 Conventional IR Methodology . . . . .	19
2.3 Review of Lifelog Benchmarking and Retrieval System . . . . .	22
2.3.1 Early Lifelog Retrieval Systems Using Multiple Techniques . .	23
2.3.2 Interactive Bechmarking of Lifelog Retrieval System . . . . .	25
2.4 Application of Computer Vision to Lifelog Retrieval . . . . .	38
2.4.1 Deep Learning . . . . .	38
2.4.2 Deep Learning in Text-Image Retrieval . . . . .	41
2.4.3 Deep Learning in Human Object Interaction Extraction . . .	42
2.5 Chapter Summary . . . . .	43
<b>3 Research Methodology and Evaluation Methods</b>	<b>44</b>
3.1 Research Methodolog . . . . .	44
3.2 Operating Constraints . . . . .	46
3.3 Evaluation Methods . . . . .	47
3.3.1 Traditional Evaluation Methods . . . . .	48
3.3.2 Enhanced Evaluation Methods . . . . .	50

<b>4</b>	<b>Datasets Collecting and Preprocessing</b>	<b>52</b>
4.1	Introduction . . . . .	52
4.2	Test Collection Construction Methodology . . . . .	52
4.2.1	Principle of Building a Lifelog Dataset . . . . .	53
4.2.2	Process of Building Personal Life Datasets . . . . .	55
4.2.3	Data Protection and Privacy . . . . .	59
4.3	Discussion & Contribution . . . . .	68
4.4	Chapter Summary . . . . .	68
<b>5</b>	<b>Baseline Search Engine for Lifelogs</b>	<b>70</b>
5.1	Introduction . . . . .	70
5.2	System1: NTCIR13-lifelog . . . . .	72
5.2.1	Search Engine Architecture . . . . .	72
5.2.2	Data Organisation and Retrieval Process . . . . .	73
5.2.3	Ranking . . . . .	75
5.2.4	Results . . . . .	75
5.2.5	Learning Improvement . . . . .	75
5.3	System2: ImageClef2017 . . . . .	76
5.3.1	Proposed Approaches . . . . .	76
5.3.2	Learning Improvement . . . . .	80
5.4	System3: LIFER . . . . .	81
5.4.1	Interactive Lifelog Search Engine . . . . .	81
5.4.2	Using LIFER for ImageCLEFlifelog2018 Retrieval Task . . . . .	84
5.4.3	Learning Improvement . . . . .	85
5.5	System4: LIFER2.0 . . . . .	86
5.5.1	Retrieval . . . . .	86
5.5.2	Experimental Configuration: LIFER 2.0 . . . . .	88
5.6	Conclusion & Contribution . . . . .	89
5.7	Chapter Summary . . . . .	91
<b>6</b>	<b>Causality for Related Query Retrieval with Deep Learned Human-Object Interactions</b>	<b>93</b>
6.1	Introduction . . . . .	93
6.2	Human Object Interaction Extraction . . . . .	97
6.3	Causality Scenario in Lifelog and Security Data . . . . .	98
6.4	Human-object Interaction Causality Model for Video and Lifelog Retrieval . . . . .	100
6.4.1	HOI Recognition . . . . .	101
6.4.2	SVO Extraction . . . . .	104
6.4.3	HOI Model . . . . .	105
6.4.4	HOI Inference . . . . .	106
6.5	HOI Annotation . . . . .	107
6.5.1	Automatic Annotation . . . . .	107
6.5.2	Manual Annotation . . . . .	108
6.6	Experimental Results . . . . .	112
6.6.1	Experimental Datasets . . . . .	112
6.6.2	Implementation Details . . . . .	113

6.6.3	Results Comparison Using Lifelog and Video Dataset . . . . .	113
6.6.4	Visualization . . . . .	116
6.7	Discussion & Contribution . . . . .	116
6.8	Chapter Summary . . . . .	118
<b>7</b>	<b>Deep Cross-Model Retrieval for Lifelog Retrieval</b>	<b>120</b>
7.1	Introduction . . . . .	120
7.2	Text Embedding for Semantic Retrieval Based on Captions . . . . .	122
7.2.1	Object Detection and Attributes Analysis . . . . .	123
7.2.2	Model Architecture . . . . .	126
7.2.3	Updated Image Captioning Generation . . . . .	126
7.2.4	Neural Embedding - BERT . . . . .	128
7.2.5	Semantic Similarity . . . . .	129
7.3	Joint Embedding for Cross-Modal Retrieval in Lifelog Retrieval . . . . .	132
7.3.1	Visual Representation . . . . .	133
7.3.2	Text-query Representation . . . . .	134
7.3.3	Unsupervised Cross-modal Common Vector Space Learning . . . . .	134
7.4	Experiment Setting . . . . .	136
7.4.1	Setting 1 - Text Embedding . . . . .	136
7.4.2	Setting 2 - VSE++ . . . . .	136
7.5	Experiments Results . . . . .	137
7.5.1	Discussion & Contribution . . . . .	139
7.6	Deployment Proposal . . . . .	142
7.7	Chapter Summary . . . . .	143
<b>8</b>	<b>Conclusion</b>	<b>144</b>
8.1	Thesis Overview . . . . .	144
8.2	Research Questions and Proposed Solutions . . . . .	146
8.3	Limitations . . . . .	148
8.4	Future Work . . . . .	148
8.4.1	Benchmarking . . . . .	148
8.4.2	Retrieval System . . . . .	150
<b>9</b>	<b>List of Publications</b>	<b>152</b>
	<b>Bibliography</b>	<b>156</b>

# List of Figures

2.1	The SenseCam viewer application . . . . .	24
2.2	Visual lifelogging “multi-axes” browser developed in 2010 . . . . .	25
2.3	liveXplore Interactive Interface from AAU . . . . .	29
2.4	UU-DCU Virtual Reality Interface showing the query and browsing mechanisms . . . . .	30
3.1	Research methodology schema. . . . .	45
3.2	Workflow of research questions and approaches . . . . .	46
4.1	A lifelogger with his everyday wearable devices: A Narrative 2 camera (wearing on his neck) and a smartwatch (the black watch on his right arm). . . . .	56
4.2	The proposed process for building a disclosed lifelog dataset. . . . .	57
4.3	An example of the XML data description for one minute . . . . .	59
4.4	The components of a published lifelog dataset. . . . .	60
4.5	The anonymisation process applied on raw lifelog data. . . . .	61
4.6	Examples of the visual anonymisation. . . . .	62
4.7	A proposed privacy-aware lifelogging model. . . . .	63
5.1	The learning improvement and updating steps of the lifelog baseline system . . . . .	72
5.2	The baseline search engine architecture . . . . .	73
5.3	Examples of the results retrieved by the proposed baseline search engine for the query “Find the moment when I was taking public transportation or taxi at sunset.” . . . . .	74
5.4	Schema of the proposed methods . . . . .	77
5.5	Schema of LIFER, the proposed interactive lifelog search engine. . . . .	83
5.6	The Interface of LIFER with an example on the results for solving a topic of "Find the moment the use is having vegetables". . . . .	83
5.7	The facets filter (left) and context browsing interface (right) of LIFER- baseline interactive search engine with an example of shopping in toyshop. . . . .	88
5.8	Result of Run 1 in various cut-off positions . . . . .	89
5.9	Result of Run 2 in various cut-off positions . . . . .	90
6.1	An example of the challenge that shows a stolen item (e.g. handbag) being passed from one individual to another: two individuals carrying the same blue backpack, with a probable handover of the backpack not captured by the surveillance cameras. . . . .	97
6.2	Scenario Map showing “dead zones”. . . . .	100
6.3	A brief illustration of our proposed model for HOI causality . . . . .	101

6.4	Image caption model . . . . .	103
6.5	HOI output example for one keyframe from video showing detected object instances and recognized the interactions between person and object they are interacting with (carry handbag). . . . .	104
6.6	Examples of HOI automatic annotation in NTCIR13 lifelog Dataset	108
6.7	HOI annotation tool . . . . .	109
6.8	The Example for Manual HOI Annotation . . . . .	110
6.9	HOI Histogram for selected manual HOI annotation . . . . .	111
6.10	HOI Histogram for selected automatic HOI annotation . . . . .	111
6.11	Result used the ICAN model and HOI-S video dataset . . . . .	116
7.1	The Bottom-up Attention model and Microsoft Computer Vision API labeled the image with objects class and corresponding attribute class, and the Microsoft Computer Vision API only focused on label image with object categories. . . . .	125
7.2	Architecture overview of the caption-based search system . . . . .	126
7.3	Architecture overview of feature extraction . . . . .	127
7.4	Baseline Architecture of Captioning Prediction . . . . .	129
7.5	Example of updated captioning results . . . . .	130
7.6	The architecture of BERT. [CLS] is the first token of every sequence.	131
7.7	The calculation of similarity . . . . .	131
7.8	Overview of Matching Procedure . . . . .	133
7.9	The transfer learning structure for a single CNN. . . . .	134
7.10	The left to VSE++ is using BERT to extract text embedding. The right to VSE++ is applying resnet152 to extract 2048 dimensions visual features. The results is calculated using the cosine-similarity .	135
7.11	Selected example of Top 5 Retrieved Results on caption based retrieval (text embedding) . . . . .	140
7.12	Selected example of Top 5 Retrieved Result using the Cross-Modal Based Retrieval (visual-semantic embedding) . . . . .	140

# List of Tables

2.1	Comparison of Information Retrieval Model . . . . .	22
3.1	Precision_Recall . . . . .	49
4.1	Statistics of NTCIR Lifelog Datasets . . . . .	64
4.2	Statistic of Lifelog Comptition and Tasks . . . . .	65
5.1	Submitted Runs. . . . .	80
5.2	Lifelog Retrieval Results. . . . .	80
5.3	Submitted Runs for LMRT. . . . .	85
5.4	Selected Criteria for LMRT Task. . . . .	85
5.5	Submitted Runs for LMRT task. . . . .	89
6.1	LSAT topics in NTCIR-13 Lifelog subtask. The causal means whether the relationship is existing with other HOIs . . . . .	95
6.2	Example of calculated Transfor Entropy scores on HOI-S dataset . .	114
6.3	Comparison results in HOI-S dataset . . . . .	114
6.4	Comparison between both automated methods . . . . .	115
6.5	Example of calculated Transfor Entropy scores on NTCIR13 lifelog dataset . . . . .	116
6.6	Overall comparison of different approaches on NTCIR13 dataset . . .	116
7.1	Comparison of Object Extraction . . . . .	124
7.2	Example of the similarity calculation between query and image captions. The query: A man is eating pasta. . . . .	132
7.3	Overall comparison of different approaches on NTCIR13 dataset . . .	138
7.4	The P@10 of 10 query sentences evaluated in the LSAT sub-task NTCIR13-Lifelog and comparison of video retrieval performance (average precision) between caption-based and visual-semantic joint embedding approaches . . . . .	139

# List of Abbreviations

<b>ANN</b>	Artificial neural network
<b>ANN</b>	Artificial Neural Network
<b>XML</b>	Extensible Markup Language
<b>NTCIR</b>	NII Test Collection for IR Systems
<b>GPS</b>	Global Positioning System
<b>TREC</b>	Text Retrieval Conference (TREC)
<b>LSC</b>	Lifelog Search Challenge
<b>ACM</b>	Association for Computing Machinery
<b>ICMR</b>	The Annual ACM International Conference on Multimedia Retrieval
<b>VR</b>	Virtual Reality
<b>LMRT</b>	Lifelog Moment Retrieval Task
<b>NLP</b>	Natural Language Processing
<b>CNN</b>	Convolutional Neural Network
<b>HOI</b>	Human-Object Interactions
<b>TF</b>	Term Frequency
<b>IDF</b>	Inverse Document Frequency
<b>LRT</b>	Lifelog Retrieval Task
<b>MAP</b>	Mean Average Precision
<b>ADLT</b>	Activities of Daily Living understanding
<b>ADL</b>	Activities of Daily Living
<b>LSTM</b>	Long Short-Term Memory
<b>GRU</b>	Gated Recur-rent Units
<b>RNN</b>	Recurrent Neural Network
<b>BERT</b>	Bidirectional Encoder Representations from Transformers
<b>LSAT</b>	Lifelog Semantic Access
<b>CV</b>	Computer Vision



# Abstract

Liting Zhou

## **Enhancing the Effectiveness of Information Retrieval Techniques for Known-Item Retrieval from Lifelogs**

Nowadays, almost everyone holds some form or other of a personal life archive. Automatically maintaining such an archive is an activity that is becoming increasingly common, however without automatic support the users will quickly be overwhelmed by the volume of data and will miss out on the potential benefits that lifelogs provide. This research is to build an effective and efficient lifelog information retrieval system, which can address the challenges of organising and searching personal life archives, using advanced IR models and ranking approaches.

The main contributions of this thesis are as follows. Firstly, lifelog data is defined and a first generation of lifelog datasets are constructed based on our proposed process. Secondly, a baseline search engine is proposed and developed for lifelog data collection, which aims to make the lifelog data searchable, organizable and serves as a baseline comparison for other lifelog retrieval systems that took part in the various benchmarking activities we organised. Thirdly, research on investigating and enriching image concept features using human-object interaction(HOI) annotations are presented and a causality score between different HOIs are calculated and automatic / manual HOI features are compared to validate which feature is most valuable for lifelog moment retrieval. Finally, the approaches presented use pre-trained textual and visual-semantic embeddings for the lifelog retrieval problem and enhance retrieval performance with a high-level semantic mapping of user information. In summary, this thesis primarily contributes to the field of building an effective and efficient lifelog information retrieval system and delivers a number of pioneering contributions.

# Chapter 1

## Introduction

### 1.1 Lifelog: Digital Diary

As technology applications become increasingly ubiquitous in our daily lives, people have begun to use various types of device to record their daily activities, i.e. cameras to record pictures and videos, driving logging devices for journey recording, and fitness trackers to record exercises [1–3] etc. Generally, the phenomenon that people digitally record common and social activities (e.g. walking, eating, travelling, drinking in the pub, etc) is referred as *lifelogging*, producing some form of a lifelog data archive. Also, the burgeoning fitness for health industry has become more extraordinarily disruptive for the general public, leading to an explosion of wearable sensors [4, 5].

*Lifelog* can be defined as a digital record of personal daily life in varying amounts of details, for a variety of purposes. The thought behind this concept is to create a permanent and searchable electronic diary of personal life experience. These personal life experience can contain information about every activity an individual participates in, such as where they go, how they get there, whom they speak with, what they see and what information they access. In effect, a multi-source, detailed, digital life diary can be generated for every individual who chooses to do so, using readily available technologies. The key-enabling technology is the availability of power-efficient sensing devices, which are embedded in the modern cell-phones and wearable devices. It is our conjecture that it is only a short step to move from sensing life to searching through the resultant personal life

archives, which is the focus of this research. We explore how state-of-the-art IR technologies can be applied to lifelog dataset.

### 1.1.1 Related Terminology

Now we should describe lifelog as it pertains to this research. Lifelogging can be described as the process by which individuals use various software or hardware devices to gather large archives of multimodal personal data from multiple sources and store them in a personal data archive, which is often referred to as a lifelog [6]. Lifelog is a generalized concept to describe recording data of daily life and the ideas behind it are much deeper and meaningful than program/data itself.

Lifelog data has been used in many domains and from prior research, we note that lifelogging tools were applied in the domains of long-term memory understanding [7], supporting human recollection [8], supporting human memory [9, 10], facilitating large-scale epidemiological studies in healthcare [11], lifestyle monitoring at the individual level [12, 13], behaviour analytics [14], diet/obesity analytics [15], or in exploring solutions for societal issues such as privacy-related concerns [16]. For many of these domains of application, lifelog data was gathered and analysed by researchers in order to draw conclusions for their research tasks.

Lifelogging is an emergent area and it also has a few new terminologies that are associated to it. For example:

- *Lifelog* is the actual data gathered. It could reside on a personal hard drive, in the cloud or in some portable storage device. The lifelog could be as simple as a collection of photos, or as large and complex as a lifetime of wearable sensory outputs (for example, GPS location logs or accelerometer activity traces);
- *Lifelogger* are the people who keep lifelogs about themselves. The first commonly known lifelogger was computer scientist Gordon Bell [17, 18], who records his life in digital form for decades as the "experimental subject" of

the project MyLifeBits [17].

- *Lifelogging* is the process of passively gathering, processing, and reflecting on the lifelog data collected by a variety of sensors, and is carried out by an individual, the lifelogger. The life experience data is mostly based on wearable sensors which directly sense activities of the person, though sometimes data from environmental sensors or other informational sensors can be incorporated into the process;
- *Egography* [19] is a process to analyse visual data captured by a wearable camera. Egocentric vision provides a unique perspective of the visual world that is inherently human-centric. Since egocentric cameras are mounted on the user (typically on the user’s head and neck), they are naturally primed to gather visual information from human everyday interactions, and can even act on that information in real-time (e.g. for a vision aid).
- *Quantified Self*<sup>1</sup> is defined to be a practice which incorporates technology into data acquisition on aspects of a person’s daily life in terms of inputs(e.g. food consumed, quality of surrounding air), states(e.g. mood, arousal, blood oxygen levels), and performance (mental and physical). It is a domain-focused effort on logging experiences (e.g. exercise levels, healthcare indicators) with a understanding of the key goals of the effort. It can be considered to be a subset of lifelog.

### 1.1.2 Lifelogging Challenge

In 1940s, Vannevar Bush introduced the world Memex, a life knowledge organisation hypermedia system operating as a desk-based device [20]. Memex introduced new concepts such as information links or trails, which are created by an individual or by others. Memex was described as an “*enlarged intimate supplement to ones memory*”. In these words, Bush identified some of the key issues for maintaining lifelogs, that

---

<sup>1</sup><https://quantifiedself.com/>

they be enlarged (store as much information as feasible), intimate (private to the owner) and supplemental (working in synergy with ones memory). Memex provided an inspiration for the Xanadu system [21] created by Ted Nelson who introduced the terms hypertext and hypermedia and was a precursor to the WWW.

In 2006, Bell and Gemmel’s work on developing the MyLifeBits [17], a personal life experience archiving tool for digital memories [18, 22], and their book *Total Recall* [22] introduced the concept of maintaining lifelogs of all information encountered and was developed using database-focused retrieval technologies.

Even multiple devices(e.g. smartphone, tablet, laptop) can be used to collect large-volume lifelog data, but in reality lifelogging has only appeared to be an extreme activity carried out only by a small number of pioneering enthusiasts e.g., Steve Mann [23], Gordon Bell [18]. The reasons why only a few pioneers have gathered these lifelogs automatically could be many and varied, and this fact also poses a set of challenges for the general community to consider:

- One of the most important challenges is the privacy and ethical concerns [24, 25]. It is our conjecture that lifelogs are likely to include the wearable cameras contents recorded by cell phone cameras, social networking sites, and location tracking services. Once the personalised experience, wellness, and memory capture and sharing provides a wide range of benefits to end users, as hard fact become clear the popularity can increase.
- A second challenge comes with the overwhelming amounts of data [26] that lifelogs naturally generate. Research has indicated that lifelog data can be summarised into useful knowledge through i) segmenting it into a series of distinct events or activities [27], ii) automatically labelling those events from both the content [28] and context [29], iii) automatically detecting faces and event novelty to identify those events that are more interesting to reflect on [30], iv) presenting segments of this lifelogs to the user as required [31] and v) analysing data to provide new knowledge or insights to the user.

- A third challenge is that lifelogs attract life activity data at various degrees of fidelity. A single sensor (e.g., an odometer or pedometer) operates at a very low fidelity whereas wearable video camera moves towards high-fidelity total capture which digitises a much richer snapshot of life activities.
- A fourth challenge was that there is no test dataset are built and released to support comparative benchmarking of lifelog retrieval systems until 2016. We addressed the challenge of how to build a disclosed lifelog dataset by proposing the principles for building and sharing such types of data. Based on these proposed principles, we will describe processes for how we built the benchmarking lifelog datasets. This is described detailedly in chapter 3.

Consequently, and in agreement with Bush’s vision of the ‘enlarged intimate supplement to ones memory’, we consider these lifelogs to be archives of sampled life experience data that can be analysed in real-time and work with the person’s memory to enhance life experience. We say they are *Personal*, in the sense that they are private archives that are gathered by the individual for the individual. *Life* in the sense that they are enlarged archives of life activity data and as such should aim for the target of total capture of life experience. *Archives* in that they will be historically stored and analysed throughout the life of the individual and potentially even longer. While real-time access provides for life enriching contextual access to past activities (in synergy with the individual’s natural memory), there is also enormous potential for long-term analysis and understanding of the life archive data.

### **Challenge of Access and Retrieval Mechanism Using Test Collections**

In order to get insights from lifelogs, individuals or researchers require tools for knowledge extraction, search, summarisation and visualisation. Moreover, to validate the performance of these tools, they need large and rich collections of lifelog data. Therefore there is a strong need to have a common published dataset for studies in this field [6]. However, the challenges in designing and construction of

a lifelog dataset that will be shared and studied in public is not trivial, arising from gathering, semantic enrichment, and pervasive accessing of these vast personal data archives [32]. To the best of our knowledge, there have been only three published lifelog datasets that fulfill the major problems above: NTCIR Lifelog [32–34] datasets.

The challenges of research with lifelog data can be summarized as follows:

- **What to log?** According to Dodge and Kitchen [35], a lifelog typically consists of numerous different types of data, such as image/video content from wearable cameras (e.g., SenseCam), audio content from personal audio devices, biometric sensor content from activity trackers (e.g., from a wristband or a phone as in [36]) or health-monitoring devices, informational content from the media consumed by a lifelogger, and so on. Ideally, we should log all information from all sources, however, it is not feasible in practice. Therefore, making decisions on what to log is indeed a non-trivial task.
- **How often to log?** Lifelog data shows considerable variance in terms of capture velocity and type variety. Some sensors, such as biometric sensors can capture data on a second-by-second basis, whereas wearable cameras may capture between 1 to 5 images per minute. In order to be useful for the lifelogger, lifelogging needs content storage and retrieval facilities that operate over data at different velocities and frequencies, in order to address a wide variety of use cases.
- **Willingness to share?** The success of the Quantified-Self (QS) movement [37] shows the willingness of individuals to gather archives of their life. People tend to be familiar with gathering their personal data, or even being a subject captured in a lifelog of someone else. However, finding people who are willing to donate months, years or even decades of data is a challenge.
- **What NOT to be shared?** Privacy and data security, which has implications for both the individual and society as a whole [38], are also important issues that need to be considered. Personally identifiable information, e.g., personal ID numbers,

car plates, addresses, and others, should not be shared. This raises the challenge of how to filter out such information in order to protect the lifeloggers as well as individuals that appear in the lifelogs.

- **Who can access the data?** The goal of a published dataset is to share with researchers. Lifelog, however, different from more traditional shared datasets [32], e.g., the privacy-aware issues, needs a well designed strategy to control data access.

We can point out that the key challenges are to gather rich archives in real-time in a non-intrusive manner, to organise these archives into meaningful experiences and generate descriptive metadata, to provide retrieval and recommendation facilities and to support omnipresence of access.

Later in this dissertation, we will describe the test collection used for this research, and also describe how they were constructed and became the dataset used by the community.

## Challenge of Information Retrieval in Lifelogging

In terms of actual functional retrieval systems for lifelog data, a number of early retrieval engines have been developed prior to current research, such as the MyLifeBits system [39], or the Sensecam Browser [27]. These systems were browsing engines, rather than search engines, and rely on a database for access. Subsequently, it was found that a faceted-multimodal search engine (even a simple one) is many times faster and more effective than browsing systems at finding known-items from large lifelogs [30], yet there were very few search engines designed for lifelog data and no means of comparing their effectiveness. This is one of our motivations for insisting on this research.

NTCIR-12 [40] is the first forum that supports a comparative evaluation of approaches to lifelog data storage and retrieval. Prior to the Lifelog task at NTCIR-12, there were no comparative benchmarking activities, and comparative and reproducible research on lifelogging was rather sparse. The main reason for



this was lack of publicly available lifelog datasets, which was due to the highly personal nature of lifelog data and the related requirements to guarantee people’s privacy when releasing such datasets for widespread use.

## 1.2 Research Motivation and Challenges

Search and ranking is only one aspect of the challenge that information retrieval systems have to tackle. When we develop retrieval systems, whether for lifelogs or simply for WWW pages, to find a starting point, we need to understand how the archives will be accessed by users. To begin our consideration, we contemplate how people access their own digital photo or video archives. For relatively small archives, a browsing mechanism (manually or automatically grouping 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, i.e. browsing a photo archive using date or location. However, when the archives become larger and less organized, a search or search/browse engine is needed to support fast and effective access.

Considering lifelogs, the size of these multi-year or multi-decade archives suggests that a browsing methodology is not sufficient from the outset. The initial experiments into multi-year multimodal lifelog search suggest that even a basic search methodology increases the possibility of a user locating desired content by a factor of three in 10th of the time [30]. When considering search, there are a number of alternative search methodologies that could be considered. Firstly, keyword based search can be used to process textual narratives generated from the sampled life experience. Another alternative approach is to support the user in generating a new type of multi-axes query in an efficient manner; for example, *I know that my friends Paul and Jack were there, it was a Sunday evening, and we were in Barcelona watching a football game in a bar.* The third option is the real-time context-driven automatic querying where real-time sampling of life

experience can trigger contextual queries to support recollection, retrieval of information and remembering intentions, which, if presented to the user in a suitable manner, can provide novel and currently unknown applications for lifelogs. Applications that can remind you that the person you have just met is having a birthday today or that the last time you bought this type of soup, you felt ill the following day.

While there has been an explosion in the amount of consumer-sensed and -generated data now being created, stored and shared, the ability to organize and provide useful retrieval facilities over this data is still limited. There are many domain-specific solutions for use, such as sensing the level of exercise or sharing a user's location, however, there are still few attempts to fully grasp the full potential of sensing the person, the quantified self [37]. As we progress towards more enhanced sensing of the person (total capture), this coming world of lifelogs will pose new challenges for the areas of multimedia contextual sensor capture, multimedia data organisation, multimedia search and retrieval as well as the human factors that define how we can interact with these lifelogs, not to mention the outstanding issues such as privacy, security of data and supporting the important human need to forget.

### **1.2.1 Capture & Storage Challenge**

The starting point in generating lifelogs is gathering data in a non-intrusive manner. The idea of 'total capture' [17], sampling life experience in high-fidelity, is that all life experience, whether considered important or useful at the time is captured which is the principle for this research. This will provide for a useful, future-proof and flexible lifelogs. As an analogy, WWW search engines index all the WWW; a search engine that selectively indexes only the important or popular content would fail to catch trends, hot-topics or the long-tail of user queries. Total capture of every aspect and moment of life experience, whether it is every heart beat, location and motion, or everything we see and do, provides us with new and potentially valuable information

about ourselves.

Inspired by the work of Gurrin et al. in [41], we list different categories of life archiving tools, which can be applied at the time of writing:

- **Passive visual capture.** Utilising wearable devices such as Narrative Clip <sup>2</sup>, the Microsoft SenseCam [42], or first-generation Augmented Reality glasses, will allow for the continuous and automatic capture of life activities as a visual sequence of digital images.
- **Passive audio capture.** Audio capture could allow for the identification of events or identification people who were speaking. It normally can be done via any smartphone.
- **Personal biometrics.** Sensing devices are becoming more common and widely used by the Quantified Self community, and they allow wearers to monitor their sleep duration, distance traveled, caloric output, and other biometrics information continually in a passive manner.
- **Mobile device context.** This refers to using a phone to continuously and passively capture users' context (e.g., location, movement, or acceleration), coupled with smart watches, these devices are able to capture much of the life activities.
- **Communication activities.** This refers to the phone or PC passively logs messages, emails, phone calls, or other contents of communications.
- **Data creation/access activities.** Logging data consumed and created, for example, words typed, web pages visited, videos watched and so on.
- **Environmental context and media.** According to [41], lifelogging is mostly, but not exclusively about recording using wearable technology, they could be logged (and accessed) by other sensors, such as surveillance cameras.

---

<sup>2</sup><http://getnarrative.com>

- Manual logging life activities. This refers to the indirect or direct logging of activity that is initiated by the user, for example, video recording, personal logs and diary.

Once sensed data is captured, it needs to be stored, which is one of the key challenges that need to be addressed. In order to support any of the 5 R's, the lifelogs should not be time-limited, i.e., should extend back indefinitely and life experiences (unless expressly requested by the user) should not be deleted for reasons of storage capacity or processing overhead.

### 1.2.2 Organisation Challenge

The human memory system has evolved over thousands of years to store autobiographical memories, and we believe that lifelogs should mimic how human memory operates. Past literature suggests that human memory stores information in distinct events or episodes, that similar episodes are associated with each other, and that more important episodes are more strongly remembered [30]. As a starting point, the lifelogs 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 unit of retrieval and choosing the basic unit is usually trivial. With lifelog data, this is not the case, since lifelogs are multimodal in nature with different modes captured at different frequencies (1 second to potentially 1 day, or longer) [25]. In order to deal with this problem, a typical approach would be to sort the data in chronological order [25], and use the minute as the basic unit of retrieval. The detailed steps for this approach will be described more detailed in chapter 3.

Building up from the basic unit, it becomes possible to organise the data at higher level which can produce more useful information systems. Typically, in a full day, we know that a person encounters anything upwards of 30 individual *episodes*

or events, with each lasting on average about 30 minutes, though there is a lot of variety [43, 44] and agreement on ‘what constitutes an event’ does not yet exist. Prior work on episode segmentation analyses sensor streams from wearable cameras to segment of life-experience into events in a pre-indexing process [27].

- *Episodes should be semantically described:* To support both post-hoc review and real-time analysis of episodes, both server-side and device-based semantic analysis tools are needed. These act as software sensors to enrich the raw sensor streams with semantically meaningful annotations. For example, raw accelerometer values on a smartphone can identify the physical activities of a user [45], bluetooth and GPS sensors or audio could facilitate the determination of where and with whom people are with [46]. Meanwhile automatic detection of concepts is possible from images [28] using deep-learning based computer vision. These user activities combined with event ontologies can help to infer higher-level semantics on the lifestyle of individuals. We currently utilise the following virtual sensors in the research in this dissertation; semantic date/time, meaningful location, personal physical activity, social interactions (via bluetooth), environmental context (via GPS and crowd-sourcing of relevant tags [47]), semantic visual concepts automatically identified from the photos and personal context of the user’s life pattern. Together, these generate a detailed description of user activities and behaviours. The specific implementation work will be expended in chapter 3.

### 1.2.3 Indexing and Annotation Challenge

To be able to retrieve life experiences for search or recommendation from lifelogs, either later or in real-time, the experiences and their annotations need to be indexed. An initial assumption would be to employ state-of-the-art techniques from artificial intelligence, database search and information retrieval to scalably index the life-experience events and provide omnipresent access via keyword/database search, ranking, recommending and presenting the multimedia

rich life experience archive through multimodal interfaces. However, we contend that to better understand how to develop lifelogging solutions and how to support effective access to these data, it becomes necessary not simply to view this as a new form of multimedia retrieval [48]challenge. Rather, it is important to understand how people will use and access their life archives. As a starting point, we turn again to the five R's of memory access from Sellen & Whittaker [26]. Each of the five Rs define a different reason why people want to access their memories, and by lifelogs. They provide valuable clues as to how to develop the organisation, search and presentation elements of lifelogs.

- Recollecting is concerned with reliving past experiences for various reasons. Recollecting will require highly accurate search engines that semantically rank content and extract just the nugget of information that is most important to the user and represent this event in as much detail as possible to aid recollection. This will require conventional information retrieval approaches to be applied to lifelog.
- Reminiscing, which is a form of recollecting, is about reliving past experiences for emotional or sentimental reasons, sometimes alone, often with others. From Information Retrieval, it will require new techniques for narrative generation [49], storytelling [50], topic detection and tracking [51] and novelty detection [52] from single (and potentially from multiple individual's archives), all operating in conjunction with conventional multimedia document ranking techniques, as required for Retrieving (below).
- Retrieving (information), is a more specific form of recollecting in which we seek to retrieve specific nuggets of information from the lifelogs. Retrieval will also require highly accurate text, multimedia and sensor data search engines. The query will define the type of knowledge that is required and in our research, we mainly focus on retrieving known-items or events to help user fully understand themselves to improve their life.

- Reflecting, is a form of quantified self analysis over the life archive data to discover knowledge and insights. It includes information summarisation from lifelog streams [53], event detection [47] and various forms of data analysis to infer and evaluate the importance of new semantic knowledge [28, 46, 54, 55] from the lifelogs. Typically such data analysis approaches rely on artificial intelligence, machine learning and various forms of statistical analysis and should proactively recommend new knowledge, not solely relying on a human information need as input, due to the fact that reflecting from lifelogs brings the potential for new knowledge discovery.
- Remembering Intentions, is a form of planning future activities which is a life activity that everyone engages in. This assists people to remind or prompt them on tasks they would like to do (*e.g. post that letter*), or real-time prompts on who they are talking to (*e.g. this is Paul*), or giving prompts on conversation cues. Past lifelogging efforts were exclusively focused on episodic memory as it was always a post-hoc analysis (i.e. constrained by technology); however now with real-time technology available we can now consider situational awareness (and past history of user) to provide prospective memory prompts.

Taking the 5Rs as a guide, it becomes possible to propose efficient methods that can effectively provide insights from the life archive, not simply in response to an explicit user query, but also in response to real-time contextual cues. Each of these five R's address different use-cases categorisations for lifelogs memory access. Once the use-cases have been defined, it then becomes important to consider the access methodologies. The challenge here is learning what the appropriate retrieval model for lifelogs and this is also the essential task in our thesis.

### 1.3 Hypothesis and Research Questions

There was no good system which is used to retrieve a lifelogger's life experience and those that exist do not provide adequate facilities [56] in 2015. There was as

no search engine, no dataset, no evaluation criteria, no standard approach to lifelog data retrieval. We have the first community benchmarking exercises [33, 40, 57–60], but they have not progressed enough yet to lead to a standard approach, as has happened with other benchmarking exercises, such as TREC. Therefore, we need to know what is an appropriate dataset and evaluation methodology, what is a basic retrieval approach that can be employed for lifelog data, how to design and enhance the baseline and finally, how to exploit the continuous and unique nature of multimodal lifelog data to enhance the retrieval process for lifelog-specific information needs. In order to find better IR model, I define the following hypothesis for my PhD research, that:

### **Hypothesis**

The application of modern multimodal information retrieval and annotation techniques enhance the performance of lifelog retrieval in a valid repeatable experimental setting.

In order to prove or disprove this hypothesis, a number of related research questions have been developed as follows:

- **Research Question 1 (RQ1). Data.**

*How to construct a test collection of lifelog data to support retrieval system comparison and how to create this test collection in a privacy-aware manner?*

Given that lifelogging is a relatively new research area, we need to define what kind of data that should be collected, and explain why and how to gather these types of data. From the collected data, propose solutions on how to organise and exploit these data in a privacy-aware and legally-compliant manner.

- **Research Question 2 (RQ2). Baseline.**

*How to apply traditional information retrieval techniques on this multi-modal lifelog data to develop a high-quality baseline search engine?*

Before 2016, we note that there was still no standard or typical lifelog



baseline search engine for benchmarking exercises. So we need such a baseline retrieval system using traditional information retrieval techniques, which can operate over the provided metadata for the collection. Additionally, the baseline search engine should be enriched, and extended with more agile and advanced solutions, which aims to give better information for higher level of insight and query engines.

- **Research Question 3 (RQ3). Ranked Model Based on Enriched HOI(Human Object Interaction) Attributes and Causality with Different HOI.**

*Can we improve retrieval performance by enriching the standard visual lifelog features by integrating causality-inspired features to provide a state-of-the-art retrieval system?*

In lifelog retrieval, there are some semantic relations, called *Causality*, that exist between different human behaviors. Causality has been viewed as a significant component to improve the accuracy of information prediction and retrieval for multimedia human-centric data [61, 62], so establishing such causal relations between variables from observation of their behaviour in time should be the basis for the construction of useful models and tools which can be used for related human behavior retrieval in lifelog retrieval.

- **Research Question 4 (RQ4). Embedding Retrieval in Lifelog Retrieval**

*Does only text embedding or cross-model joint embedding enhance the retrieval performance beyond the state-of-the-art techniques and if so, by how much?*

Lifelog data search has attracted great interest, but how to rank search results based on relevance level is always challenging area, especially with visual content of images and tags. In our research, we simultaneously investigate text and visual representations of images and accomplish deep learning approaches to estimate relevance.

## 1.4 Research Contributions

In this section, the key contributions made in this thesis are listed:

- Identified the challenges for building a shared lifelog dataset;
- Proposed principles for building a shared lifelog dataset.
- Based on the proposed principles, built and described the whole process from data gathering to determine the roles for the people who are building, sharing and exploiting a disclosed lifelog dataset, and released the first dataset for widespread use.
- Proposed how future lifelog data should be stored and organised, and therefore made easy for a user to access via an appropriate search mechanism;
- Defined a high-quality baseline search engine using standard IR approach;
- Integrated potential causality between different HOIs into lifelog retrieval process and highlighted its effectiveness;
- Integrated visual-semantic embedding to enhance the retrieval performance.

## 1.5 Thesis Outline

This thesis is mainly focused on exploring personal lifelogging know-item retrieval and applying new state-of-the-art techniques to make lifelog more accessible in terms of data collection, organization and visualization. In this chapter, we introduced the motivation, challenge and the significance of this research. We also proposed our research questions and summarised our research contributions. The remainder of this thesis is organized as follows:

- Chapter 2 outlines currently existing work that relates to lifelog moment retrieval tasks. We discuss the conventional information retrieval methodology, review the lifelog benchmarking and the retrieval system from participants, and explain the difference with our system.

- Chapter 3 presented research methodology and answer the four research questions and the evaluation methods employed in this research are also covered in this chapter.
- Chapter 4 presented and discussed how to construct three lifelog dataset based on the process including data collection, data organising and data publishing.
- Chapter 5 motivates the need for, and presents an improved baseline search engine for lifelogs, which is used to provide a starting point for researchers in the area, as well as a documented system for comparative analysis.
- Chapter 6 further improves the retrieval performance through enriching the HOI features and the possible causality between different HOIs.
- Chapter 7 presents our work on applying the embedding techniques to improve the performance in lifelog moment retrieval.
- Chapter 8 summarises the work of this thesis, discusses some recent developments and explores a few directions for future research.

## Chapter 2

# Related Work and Background

### 2.1 Introduction

Based on our research questions, it is obvious that a good IR methodology can make a lifelog retrieval system more efficient and effective. So firstly this chapter provides a comprehensive review of conventional IR methodologies applied in lifelog retrieval literature which mainly focuses on a critical analysis of existing approaches to lifelog information retrieval. Then this chapter serves as a detailed introduction to lifelog bechmarking challenge, which forms the basis for the fields we are contributing to in this thesis.

### 2.2 Conventional IR Methodology

Lifelog is a media rich repository and the access mechanisms must support a large number of use-cases. A good retrieval system will be necessary for retrieving information that is useful to a user for any given information need, or that can be used to underpin many applications, as shown in [17, 18]. An initial assumption would be to employ state-of-the-art techniques from database search to index the life-experience data in lifelog and provide access via keyword/database search, ranking though a desktop interface. This is the approach taken by the MyLifeBits project at Microsoft Research. As lifelog archives grow larger, the set-based retrieval model (as used in relational databases) very quickly becomes unworkable. In a similar manner to the early progress of IR technologies, the move towards

ranked output for ad-hoc queries becomes necessary. In order to improve the effectiveness and efficiency of lifelog retrieval, we need IR technologies which can be applied into our lifelog data. So here we briefly review conventional IR methodology. To make information retrieval efficient, conventional text documents are normally transformed into a suitable representation. Now such type of information is retrieved efficiently with the help of IR models. The traditional models are categorized into similarity-based models and probabilistic models. The similarity-based model represents documents as sets of words or phrases. Similarities are usually derived from set theoretic operations on those sets. Boolean model, Vector Space Model, Latent Semantic Indexing are all exemplified. As to the Probabilistic models, BM25 and language models for IR belongs to this group. Also all these models are categorised as traditional models.

### **Boolean Model**

The Boolean Model of information retrieval is a classical first model and considered to be an exact-matching model. The Boolean Information Retrieval is based on Boolean logic and classical set theory in that both the documents to be searched and the user's query are conceived as sets of terms. So in Boolean model, documents are associated with a set of keywords and for query formulation operators like AND, OR, NOT are used [63]. The search engine only returns documents that satisfy the Boolean expression, so the retrieval function of this model treats a document as either relevant or irrelevant. This is similar to how the MyLifeBits system worked and given a large archive of lifelog data, this concept of non-ranked retrieval would not work. A ranked list is retrieved.

### **Vector Space Model (VSM)**

VSM is a simple model based on linear algebra. The procedure of this model is divided into three stages: Document indexing (content bearing terms are extracted from the document text), weighting the indexed terms (enhancing the retrieval of

documents is relevant to user) and ranking the documents (by similarity measure). A common similarity measure is known as cosine measure determines angle between the document vector and the query vector [64]. When the user requests for some information, the output are generated based on the similarity between the query vector and the document vector.

**TF-IDF** is normally a classical vector space model and stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. The *term frequency* of a word in a document. There are several ways of calculating this frequency, with the simplest being a raw count of instances a word appears in a document. Then, there are ways to adjust the frequency, by length of a document, or by the raw frequency of the most frequent word in a document. The *inverse document frequency* of the word across a set of documents. This means, how common or rare a word is in the entire document set. The closer it is to 0, the more common a word is. This metric can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm.

## Probabilistic Model

The probabilistic retrieval model is based on the Probability Ranking Principle, which states that an information retrieval system is supposed to rank the documents based on their probability of relevance to the query. Documents and queries are represented by binary vectors  $d$  and  $q$ , each vector element indicating whether a document attribute or term occurs in the document or query or not [65]. Instead of probabilities, the probabilistic model uses odds  $O(R)$ , where  $O(R) = P(R)/1 - P(R)$ ,  $R$  means “document is relevant” and  $-R$  means “document is not relevant”. BM25 belongs to probabilistic model.

**BM25** stands for “Best Match 25”. It is the 25th iteration of tweaking the relevance computation. BM25 has its roots in probabilistic information retrieval. Probabilistic information retrieval is a fascinating field unto itself. Basically, it casts

relevance as a probability problem. A relevance score, according to probabilistic information retrieval, ought to reflect the probability a user will consider the result relevant.

Table 2.1: Comparison of Information Retrieval Model

Attributes/Models	Boolean	Vector space	Probabilistic
Advantages	Easy to implement	Simple model, weights are not in binary	Theoretical adequacy: ranks by probabilities
Disadvantages	Does not rank documents, retrieves too many or too few	Suffers from synonymy and polysemy. It theoretically assumes that terms are statistically independent	Binary weights, ignore frequencies and independence assumption

## 2.3 Review of Lifelog Benchmarking and Retrieval System

Lifelogging typically occurs in a passive manner (i.e., using sensors and not relying on human input). A commonly used definition of lifelogging is as ‘*a form of pervasive computing, consisting of a unified digital record of the totality of an individual’s experiences, captured multimodally through digital sensors and stored permanently as a personal multimedia archive*’ [35]. Lifelogging can generate enormous (potentially multi-decade) archives that are too large for manual organisation. What sets lifelogging apart from conventional personal data organisation challenges (e.g., photos or emails) is the fact that lifelogs, being captured passively, are typically continuous in nature and are non-curated archives. Hence these lifelogs pose a significant challenge for researchers to develop appropriate information organisation and retrieval approaches.

Currently, there are numerous data organisation and retrieval systems designed for lifelog data, in this discussion we focus on interactive systems (i.e. more than

query/submit pairs) for multimodal lifelog data archives. The seminal MyLifeBits [17] project at Microsoft produced, what is generally regarded as the first interactive lifelog retrieval system, which was based on a database indexing and retrieval metaphor. Lee et al. [66] went beyond the database metaphor by developing an interactive event-organised lifelog browsing interface for visual lifelog data that segmented days into events, based on analysis of visual and sensor data, and linked events together in a single diary-style interface. More recently, the LEMoRe [67] system, an interactive lifelog retrieval engine, developed in the context of the Lifelog Semantic Access Task (LSAT) of the the NTCIR-12 challenge, integrated classical image descriptors with high-level semantic concepts and was powered by a graphical user interface that uses natural language processing to process a user's query.

### 2.3.1 Early Lifelog Retrieval Systems Using Multiple Techniques

In relation to standalone retrieval efforts, early research on lifelog retrieval has focused on using *images* as unit of retrieval (e.g., [66]) with some early work in supporting user browsing these image collections [27], or on the use of maps metadata, such as GPS locations, to organise content visually [68]. We refer the reader to [6] for an overview of early efforts at lifelog search and retrieval. Significant efforts also went into the development of graphical user interfaces to visualize the data and also provide a positive user experience.

Most of lifelogging research has focused predominantly on visual lifelogging storing and retrieval model in order to capture life details of life activities, we review the early lifelog retrieval system as below:

- **MyLifeBites(2002)** MyLifeBites is a browsing engine for storing all of personal digital data, including documents, images, sounds, and videos. It mainly relied on state-of-the-art data search techniques to scalably index the life-experience events in lifelog and support access via keyword/database search, ranking though a web interface. It employed three designed tables:



one for media data, one for annotation and links and one for collection links. The search can be made for resource descriptions, blobs and specified annotation text, which had been lined to the source of the annotation.

- **SenseCam: A Retrospective Memory Aid(2006)- Data-based Access**

Sensecam [27] was a tool which is designed to capture a digital record of the wearer's life. The SenseCam captures a series of still images based on movement and changes in the intensity of light and heat and support the recollection of everyday experience, as well as retrieval of information about past events. However, the same study showed that the capacity for these images to help people recollect their past experience rapidly decreased after only three months, casting doubt on whether such devices can support longer-term recollection. In addition to sensecam device, it also offered a viewer application(shown in Figure 2.1) for managing and reviewing image sequences captured by the device at any captured date.

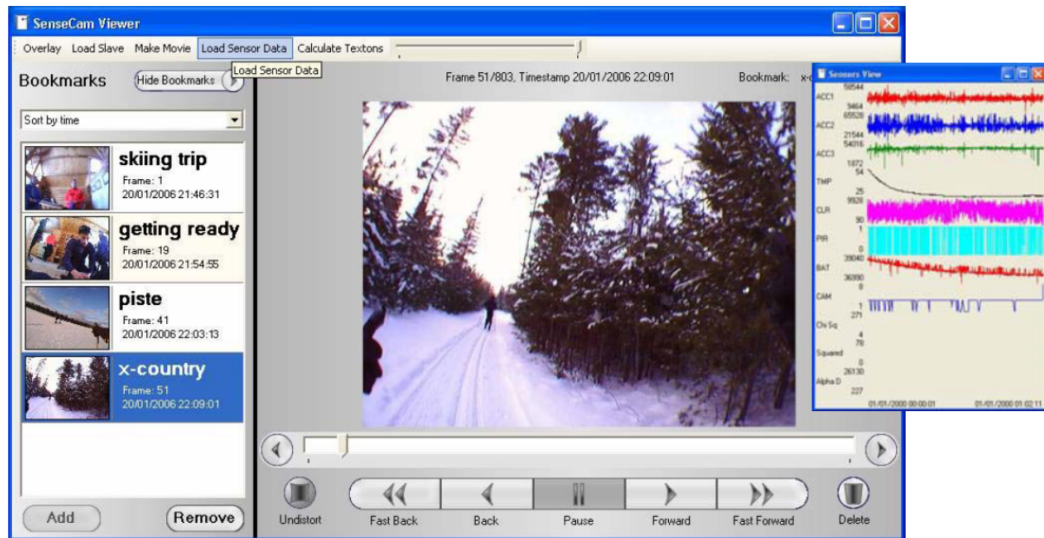


Figure 2.1: The SenseCam viewer application

- **Faceted Retrieval System(2008)**

Based on previous work on early retrieval engines, it was found that a faceted-multimodal search engine (even a simple one) was many times faster and more

effective than browsing systems at finding known-items from large lifelogs [69], yet there were few search engines designed for lifelog data and no means of comparing their effectiveness. The first faceted retrieval system for lifelog data over large archives was developed by Lee et al [66](2008); Doherty et al [69](2010), which is showed in Figure 2.2. This lifelog search engine employed a photo search metaphor of event segmentation, event annotation and multi-axes search. Not surprisingly, it was found that the search metaphor was significantly more useful than a time/date browsing metaphor.

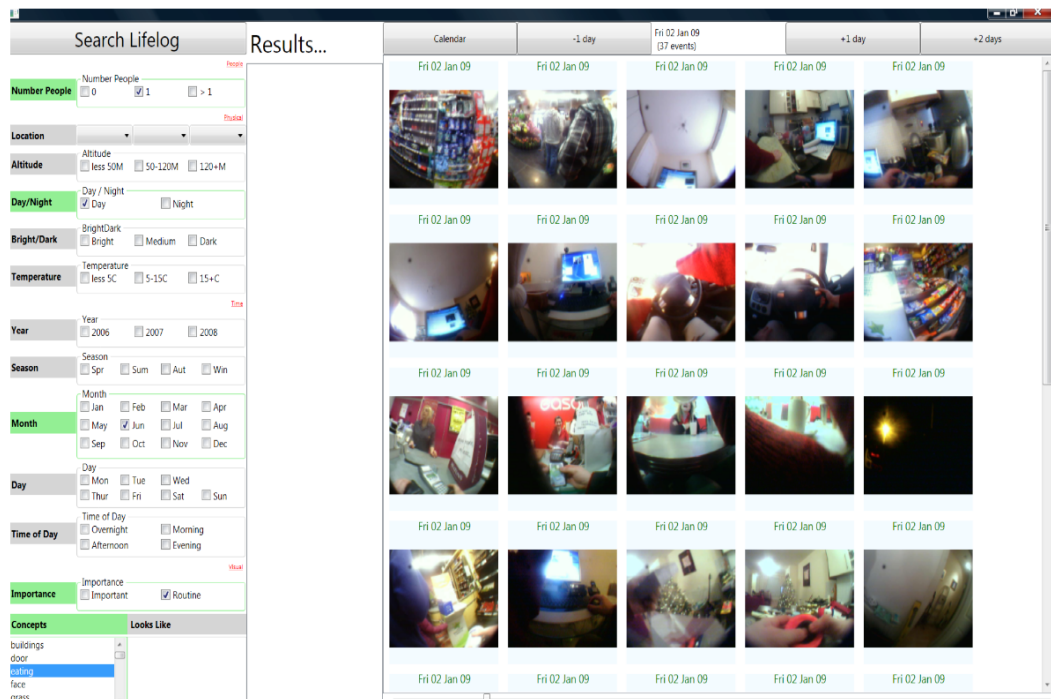


Figure 2.2: Visual lifelogging “multi-axes” browser developed in 2010

However, there is no comparative analysis, so we need benchmarking for encouraging more researchers participants into lifelog research.

### 2.3.2 Interactive Benchmarking of Lifelog Retrieval System

The field of information retrieval has a long history of benchmarking exercises in which numerous systems and techniques to solve specific retrieval challenges are compared against each other by using the same test collections openly and

cooperatively. Typically this works by participants developing systems, evaluating them over test collections and then coming together for an open comparison of system performance. This is best exemplified by the test collection methodology employed by large-scale international efforts, such as TREC [70], CLEF [71], NTCIR [72] and in the multimedia field, efforts such as ImageCLEF [73] or MediaEval [74]. A summary of these activities and their challenges can be found at [75].

The lifelog field has seen a number of such initiatives. The current Lifelog competition was a comparative benchmarking exercise with the aim of encouraging research into the organisation and retrieval of data from multimodal lifelogs. As an example NTCIR-lifelog supported participants to submit to five subtasks, each tackling a different challenge related to lifelog retrieval. Many good examples of interactive interfaces can be seen in the systems developed for the interactive lifelog search tasks in Lifelog competition. Here we list the three interactive benchmarking competition of lifelog retrieval and the system designed by the corresponding participant teams. During my phd, I was a founder of these tasks.

## **NTCIR**

The Lifelog task at NTCIR was a comparative benchmarking exercise with the aim of encouraging research into the organisation and retrieval of data from multimodal lifelogs. The Lifelog task ran for over four years from NTCIR-12 until NTCIR-14 (2015.02-2019.06) and it aims to advance the state-of-the-art research in lifelogging as an application of information retrieval.

As our research mainly focuses on the lifelog moment retrieval task, so we will lay emphasis on LSAT (lifelog semantic access task) task in NTCIR-lifelog. Over the three instances of the LSAT Task, we note that task participants took many different approaches to the development of retrieval systems. Given that there are no standardised baselines that can be applied, this is not surprising. Participating teams developed many different experimental systems, both interactive and

automatic in nature. We look firstly at interactive retrieval engines over the three editions of NTCIR-Lifelog. At NTCIR-12, the participating team from University of Barcelona (Spain) developed the only interactive retrieval engine that integrated a semantic-content tagging tool to enhance the quality of the annotations [67]. At NTCIR-13, the DCU team (my work) employed a human-in-the-loop to translate the provided queries into system queries for their retrieval engine, in one of their runs [76]. However, at NTCIR-14, we note that three of the participants developed interactive systems and a fourth participant also integrated the human-in-the-loop query enhancement. NTU (Taiwan) developed an interactive lifelog retrieval system that automatically suggested to the user a list of candidate query words and adopted a probabilistic relevance-based ranking function for retrieval [77]. They enhanced the official concept annotations and pre-processed the visual content to remove poor quality images and to offset the fish-eye nature of the wearable camera data. DCU (my work) developed an interactive retrieval engine for lifelog data [78] that was designed for novice users and relied on an extensive list of facet filters over provided metadata. Finally, the VNU-HCM (Vietnam) group developed an interactive retrieval system [79] that used enhanced metadata and visual enrichment, sometimes including human annotations. Their scalable and user-friendly interface to this system significantly outperformed competing systems at NTCIR-14, due primarily to the enhanced annotations. As expected, all interactive runs significantly outperformed the automatic runs at each edition of NTCIR-Lifelog [33, 40].

### **LSC-challenge**

The Lifelog Search Challenge (LSC) at annual the ICMR conference is an international content retrieval competition that evaluates search for personal lifelog data. At the LSC, content-based search is performed over a multi-modal dataset, continuously recorded by a lifelogger over 27 days, consisting of multimedia content, biometric data, human activity data, and information activities data. In

this work, the first LSC were reported and took place in Yokohama, Japan in 2018 as a special workshop at ACM International Conference on Multimedia Retrieval 2018 (ICMR 2018). The LSC competition highlight advances in the state-of-the-art for interactive lifelog retrieval by collating in 2018, six interactive retrieval systems developed for the first collaborative benchmarking exercise for lifelog information retrieval (LSC 2018), 12 teams took part in 2018 and 14 in 2020. These teams had all indexed the dataset prior to attending the workshop and then during the interactive search challenge, both expert and novice users took part in evaluating the performance of the six systems. For the challenge, each participant was given a desk with a clear view of a large screen which showed the topics, the time remaining on each topic, as well as the current and overall scores of each team. Here we briefly introduce and discuss the work from notable participating teams.

- **AAU(liveXplore)** The successful employment of the web technologies-based diveXplore system [80] by Alpen-Adria-Universität Klagenfurt (AAU) at past iterations of the annual Video Browser Showdown led to the development of liveXplore [81], a system modification serving as a lifelogging data browser by focusing on visual exploration and retrieval as well as metadata filtering. Since the application is developed for processing video scenes, LSC image sequences were converted to video using a constant frame rate. Pre-calculated semantic shot segmentation enabled clustering of similar images to coherent scenes and the creation of the main interface, which presented the user with an adjustable multi-level feature map grouping together similar shots according to machine learning descriptors or handcrafted features. Additionally to providing shot-specific similarity search based on these features, liveXplore specifically offered the possibility of exploring individual lifelog day summaries as chronologically ordered galleries as well as videos in an overlay view enriched with metadata information. Finally, in order to search the data according to metadata information the system featured a filter view that allowed users to mix and match temporal,

location or activity-based and machine learning concept oriented filtering.

The liveXplore interface is shown in Figure 2.3.



Figure 2.3: liveXplore Interactive Interface from AAU

- **The Interactive Lifelog Image Browser** developed by Universitat Politcnica de Catalunya (UPC) in collaboration with Dublin City University (DCU), was a novel retrieval engine based on three core considerations: (1) the development of a multi-faceted query interface, (2) the inclusion of a trusted retrieval engine, and (3) the novel presentation of a ranked list of results [82]. Borrowing from the standard WWW-interface for faceted search systems (e.g. hotel booking or flight booking), the interface was designed with two sections. On the left side the query panel is displayed which contains the faceted and free-text query elements. On the right side is the result display panel. The faceted search components included Day-of-the-Week selector, Calendar selector, moment-of-the-day selector (time of day), Place selector and Heart-rate. The ranking engine indexed every minute as the retrievable unit using the commonly used TF-IDF ranking methodology. In order to provide the user with some context of a ranked moment, the previous two images and the following two images contribute (on a sliding scale) to the

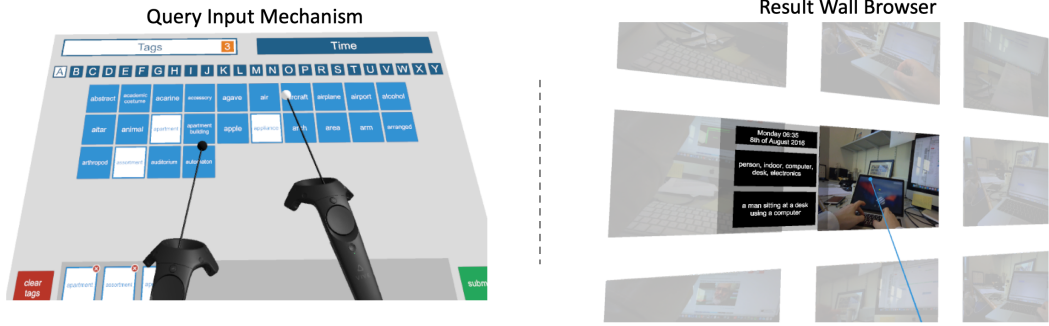


Figure 2.4: UU-DCU Virtual Reality Interface showing the query and browsing mechanisms

overall score of the main image. Selecting an image allows it to be submitted to the server for judgement.

- The Virtual Reality Lifelog Explorer** explorer developed for the LSC in a collaboration between University of Utrecht (UU) and Dublin City University (DCU) has two components, each of which needed to be optimised for a VR environment [83]. The querying component was a virtual interface designed to provide a quick and efficient means for a user to generate a filter query within the VR system. This gesture-based querying interface consisted of two sub-menus, one for selecting lifelog concepts of interest and the second for selecting the temporal aspect of the query (e.g. hours of the day or days of the week). The ranking was based on a combination of concept relevance and the time of capture (maintaining the temporal organisation of the data), where concept relevance took precedence over the temporal arrangement. Any image displayed on the VR ranked list could be selected for further exploration by pointing the user's controller at it and pressing a button (see Figure 2.4, right-side). This showed additional metadata about the image such as the specific capture date and time and what concepts have been detected.
- VNU-HCM(Semantic Concepts Fusion Retrieval)** from the University of Science and University of Information Technology (Vietnam National

University-Ho Chi Minh city) developed a pioneering lifelog retrieval system that integrated recent achievements in computer vision for place and scene attribute analysis, object detection and localization, and activity detection using image captioning [84, 85]. This system can be highlighted according to the three main novel advancements: (1) Visual Clustering for Images: independent images are organised into visual shots, sequences of similar images, based on visual information, then visually similar sequences are linked to a scene using visual retrieval with Bag-of-Word framework, (2) Concept Extraction: the system extracts the location of as well as the scene attributes of an image and create a textual caption of the image for indexing, (3) Augmented Data Processing: besides visual information, lifelogging data also contain useful augmented data, such as biometrics, blood pressure, blood sugar level, text data of computer activities, etc. Indices were created for such augmented data in an indexing process. The overall interface allowed the user to integrate all of these core techniques in one comprehensive system, with the query panel on the left and the result panel on the right.

- **The SIRET(VIRET - An Interactive Lifelog Search Engine)** team from Charles University, Prague, participated also at the Lifelog search challenge with an updated version of the VIRET system [86]. The objective of the participation was to inspect the performance of a purely content-based video retrieval tool for Lifelog data. The tool did not consider provided lifelog specific modalities (e.g., locations or heart rate). Since the tool relies on sequences of extracted video frames, the transition to the visual Lifelog repository was straightforward. Every day from the collection was treated as one ‘video’ represented by the lifelog images, extended by selected images/frames extracted from provided short videos. For each image, automatic annotations were obtained from a retrained GoogleNet (with an own set of 1,390 ImageNet labels). In addition, a colour signature for sketch-based search and deep feature vector from the original GoogleNet were



extracted. Based on the automatically extracted features, users could provide three types of query input (keywords, colour sketch and example images) that could be further combined by a late fusion strategy. Even though the tool performed relatively well (the overall third place), it turned out that the additional lifelog modalities would be important for effective filtering. Therefore, they plan to incorporate the modalities in the future versions of the VIRET tool.

**Comparison:** Table 2.2 shows a basic comparison between features implemented in each of the system in LSC-challenge workshop. Some features were expected to provide obvious utility to developers, such as the facet filters which were employed in some form by all systems. Most systems also incorporated some form of event/scene organisation in the user interface, as well as producing a novel form of ranked list in response to a user information need. Interestingly, only half of the systems actually implemented biometric filters as part of the query process. Finally, we note that two of the systems (liveXplore and VIRET - two of the top three ranked systems) were based on existing video browsing/retrieval systems, which were refined to work with lifelog data.

Table 2.2: Summary of the features used by all six participating systems. "DCU" team is based on our LIFER system

Feature	AAU	UPC-DCU	UU-DCU	VNUHCM-US	VIRET
Facet Filters	Y	Y	Y	Y	Y
Event/Scene Organisation	Y	N	N	Y	Y
Visual Clustering	Y	N	N	Y	N
Novel Ranked List Visualisation	Y	Y	Y	N	Y
Enhanced Visual Analytics	Y	N	N	Y	Y
Integration of Biometric Data	Y	N	N	Y	N
Non-textual/faceted Querying Mechanism	Y	N	N	Y	Y
Based on Existing Video Search Tool	Y	N	N	N	Y

## ImageClef

Since 2016 with the first lifelog initiative, the NTCIR-12 - Lifelog task, research in lifelogging, there was a need new ways of organising, annotating, indexing and interacting with lifelog data, and that is the key motivation of the Lifelog task at ImageCLEF, which I confounded. The ImageCLEFlifeLog tasks at

ImageCLEF2020, ImageCLEF2019, ImageCLEF2018, ImageCLEF2017, which were inspired by the fundamental image annotation and retrieval tasks of ImageCLEF. Since 2017, the task continued to follow the general evolution of ImageCLEF, by applying the advanced deep learning methods and extending the focus to multi-modal approaches instead of only working just with image retrieval. There are two sub-tasks (challenges) each year in ImageCLEF-lifelog task: Activities of Daily Living understanding (ADLT) (later changed to Solve My Life Puzzle, which focused on the new ways of organising lifelog data, in particular in rearranging lifelog moments) and Lifelog Moment Retrieval (LMRT). As our research mainly focuses on the lifelog moment retrieval task, so we will lay emphasis on LMRT task. Over the three instances of the LMRT Task, we note that task participants took many different approaches to the development of retrieval systems.

At Imageclef-17, in total the three groups that took part in the task and submitted overall 19 runs. All three participating groups submitted a working paper describing their system, thus for these there were specific details available:

- I2R [87]: This team got the best F1@10 measure score(excluding the organizers' team) of 0.497 by building a multi-step approach. As first step they filtered out uninformative images. Then the system ranked the remaining images and clustered the top ranked images into a series of events using either k-means or a hierarchical tree. As final step they selected, in an iterative manner, as many images per cluster as to fill a size budget.
- UPB [88]: This team proposed an approach that combines textual and visual information in the process of selecting the best candidates for the tasks requirements. The algorithm starts by analyzing the concept detectors output provided by the organizers and selecting for each image only the most probable concepts. From the list of the topics, each of them has been then parsed such that only relevant words have been kept and information regarding location, activity and the targeted user are extracted as well. The

images that did not fit the topic requirements have been removed and this shortlist of images is then subject to a clustering step. Finally, the results are pruned with the help of a similarity scores computed using WordNets builtin similarity distance functions.

At Imageclef-18, the number of participants was considerably higher with respect to 2017: there are in total 41 runs: 29 (21 official, 8 additional) for LMRT sub-task, from 5 teams from Brunei, Taiwan, Vietnam, Greece-Spain, Tunisia, Romania, and a multi-nation team from Ireland, Italy, Austria, and Norway.

- CAMPUS-UPB [89] focused on LMRT. In their methods they analysed visual information, textual information and metadata. Visual concepts are extracted using a convolutional neural network (CNN) approach. Visual features are then clustered using K-means and reranked using the concepts and queried topics.
- AILab-GTI [90] proposed a weakly supervised learning method. The method consists of three different strategies. The Two-class strategy, is based on deep learning and presents each topic by two classes one described by the topic and the other by the absence of it. The second strategy, Ten-class strategy, considers all classes at the same time. The final strategy, called Eleven-Class strategy is similar to the previous one with one additional class for topics not belonging the the challenge.
- The NLP-Lab [91] team tackled both subtask of the ImageCLEFloggin task. The main idea was to reduce user involvement during the retrieval by using natural language processing. For both tasks specific approaches were presented based on the same methodology. Visual concepts are extracted from the images and combined with textual knowledge to get rid of the noise. For LMRT task, ranking is performed exploiting similarity between image concepts and user queries.

- HCMUS [92] proposed a method based on visual concept fusion and text-based query expansion for both sub tasks. First concepts are extracted from the images. In addition textual descriptions of the images are created. These information are then combined in an inverted index for retrieval. To determine the similarity between words and phrases word embedding is used. Based on this and the users provided queries semantically similar concepts are recommended to the users.
- The Regim Lab [93] team decided to work on the LMRT task. Combinations of visual features, textual features and a combination of both were used. For the visual features fine tuned CNN architectures were utilized. For the combination of visual and textual features the best visual run was combined with XQuery FLOWR results.

At Imageclef-19, the number of participants as well as the number of submissions was considerably higher with respect to 2018: there are in total 50 valid submissions (46 official and 4 additional) for LMRT, from 10 teams representing over 10 countries.

- The REGIM-Lab [94] focused on LMRT task by improving the system from their last year participation with NoSQL which offers distributed database and framework to handle huge data. They employed the ground-truth of development set to improve the fine-tuning phase for concept extraction. In addition, CQL Query was used to exploit complicated metadata. For the query analysis, the authors trained a LSTM classifier to enhance the query with relevant concepts.
- The UPB team [95] proposed the algorithm to eliminate blurry images which contain less information using a blur detection system. Following that, a metadata restriction filter, which was created manually by users, was applied to the dataset to further remove uninformative images. The remaining images was then computed a relevance score based on given metadata description for query answering.

- The UAPT Bioinformatics (UAPT) team [96] proposed an automatic approach for LMRT task. The images are pre-processed through an automatic selection step to eliminate images with irrelevant information to the topics (feature extraction, machine learning algorithm, k-nearest neighbors, etc.) and more visual concepts were generated using various state-of-the-art models (Google Cloud Vision API, YOLOv3). Then they extracted relevant words from topics' titles and narratives, dividing them into five categories, and finally matching them with the annotation concepts of lifelog images using a word embedding model trained on Google News dataset. Moreover, an extra step to reuse unselected images from the pre-processing steps for image similarity matching was proposed to increase the performance of their system.
- The TUC\_MI team [97] proposed an automatic approach for LMRT task. They firstly extracted twelve types of concept from different pre-trained models to increase the annotation information for lifelog data (1191 labels in total). For image processing, two methods were introduced to transform images into vectors: image-based vectors and segment-based vectors. For query processing, they processed the query with Natural Language Processing techniques and introduced a token vector which has the same dimension as image/segment vector. Finally, they defined a formula to compare the similarity between image/segment and token vector and conducted an ablation study to find the best model that achieved the highest score in this task.
- The HCMUS team [98] proposed to extract semantic concepts from images to adapt to lifelogger's habits and behaviours in daily life. They firstly identified a list of concepts manually, then trained object detectors to extract extra visual concepts automatically. Moreover, they also utilised object's region of interest to infer its color by K-Means clustering. To further understand the temporal relationship between events, they also integrated the visualization function for an event sequence in their retrieval system.

- The BIDAL team [99] participated in LMRT tasks. For the LMRT task, they introduced their interactive system with two main stages. For stage 1, they generated many atomic clusters from the dataset based on rough concepts and utilising text annotation to create Bag-of-Words (BOW) vectors for each image. In stage 2, they generated the BOW vector for query texts and found similar images that suited the context and content of the query. They then used the output for result expansion by adding more images which were in the same cluster. Finally, an end-user chooses appropriate images for the query.
- The ZJUTCVR team [100] pre-processed the images with blur/cover filters to eliminate the blurred and occluded images. Then, they proposed three approaches to handle the remaining lifelog images: the two-class approach, the eleven-class approach, and the clustering approach. For two-class approach, the authors divided query topics into directories and ran a test on each directory with a fine-tuned CNN. After that, the results are classified into two classes based on the relevance to topic description. The eleven-class approach shared the same process with the previous method, but the results are split into 11 classes, where 10 classes are corresponding to 10 query topics and the 11th class contains irrelevant images to all 10 topics. With the clustering approach, the team inherited the procedure of two-class approach with a modification after the first-round retrieval by clustering images with LVQ algorithm.
- The ATS team [101] approached the LMRT task with 11 automatics runs and 1 interactive run. All automatic runs shared the same process with 4 components: Interpretation, Subset selection, Scoring and Refinement, but differed in the configuration of selecting the approach of each component. The interpretation state provided keywords and synonyms approaches which utilised WordNet and Word2Vec to diversify the results. The choice of subset is highly reliant on the configuration to use partial match or entire dataset to test. A scoring process was used to produce final ranking with three settings:

label counting, topic similarity and SVM. The Refinement step offered multiple approaches: weighting, thresholding, visual clustering and temporal clustering. Finally, the team conducted ablation study to find the best configuration. The interactive run was done by letting user filter the subset of dataset and choose automatic approach of each component to complete the query.

Over the three instances in ImageCLEF lifelog task, the participants took many different approaches to the development of retrieval systems. In Imageclef-17 and Imageclef-18, many participants mainly used more semi-automatic approaches, which combine human knowledge with state-of-the-art multi-modal information retrieval. In Imageclef-19, most participants utilized directly tailored-built deep networks or exploiting the semantic concepts extracted by using deep learning methods.

## **2.4 Application of Computer Vision to Lifelog Retrieval**

### **2.4.1 Deep Learning**

Traditional machine learning algorithms including Decision Tree, Random Forest, SVM etc. include one layer of learning and require heavy feature engineering before the data gets fed into algorithms, however, deep learning utilises machine learning algorithms that learn features automatically using multi-layer networks with non-linear processing units in each layer [102]. Multiple levels of learned representations in deep learning lead to various levels of abstraction; the levels corresponding to hierarchical latent features and higher-level features that are obtained from lower-level features [102]. Deep learning methods encompass deep neural networks(DNN), hierarchical probabilistic models, and various unsupervised and supervised feature learning algorithms.

State-of-the-art deep learning architectures include multi-layer deep neural networks, Recurrent Neural Networks (RNNs), Long Short-Term Memory

(LSTM) [103]/Gated Recurrent Unit (GRU) [104], Deep Belief Networks (DBN) [105], Deep Stacking Networks (DSNs) [106], Convolutional Neural Networks (CNNs) [107], Bidirectional Encoder Representations from Transformers (BERT) and more. CNNs in particular have been tremendously successful in many real-world applications. In this thesis, we mainly use CNNs and LSTM for learning visual features and BERT for learning text features.

#### **2.4.1.1 Feature Learning**

Feature learning refers to techniques that learn to transform raw data input to an effective representation for further higher-level processing such as classification, automatic detection, and segmentation. Feature learning approaches provide a natural way to capture cues by using a large number of code words (sparse coding) or neurons (deep networks), while traditional computer vision features, designed for basic-level category recognition, may eliminate many useful cues during feature extraction [108]. Deep neural networks are multi-layered and they are used to learn feature representations in the hidden layer(s). These representations are subsequently used for classification or regression at the output layer, and feature learning is an integral part of deep learning [109].

#### **Text Feature Representation in NLP**

The process of transforming text into numeric features, is usually performed by building a language model. These models typically assign probabilities, frequencies or some calculated numbers to words, sequences of words, group of words, section of documents or whole documents. The most common techniques are: One-hot encoding, N-grams, Bag-of-words, vector semantics (TF-IDF), distributional semantics (Word2vec, GloVe). Recently, very powerful language models have been developed such as BERT built by Google. BERT is Google's neural network-based technique for natural language processing (NLP) pre-training. In this thesis, One-hot Encoding and BERT are the two main language models we used for



textual embedding transformation.

- **One-hot Encoding:** It is the classic approach to encode categorical variables for statistical analysis. It has been applied into some domains of computer vision [110], Data analysis [111]. However, the one-hot encoding leads to feature vectors of high dimensionality when applying it to high-cardinality data. This is especially problematic in big data settings, which can lead to a very large number of categories, posing computational and statistical problems.
- **BERT:** BERT (Bidirectional Encoder Representations from Transformers) is a language model, recently published by researchers at Google AI Language. It has attracted large attention in the Machine Learning community by presenting state-of-the-art results in a wide variety of NLP tasks, including Information Retrieval [112], Question Answering [113], Natural Language Inference [114], grounding referring expressions [115]. BERT's key technical innovation lies in applying the bidirectional training of Transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. BERT is the most popular technique in NLP domain currently. More and more researchers have started to feed individual sentences into BERT and to derive fixed-size sentence embeddings. Most researchers mainly focused on fine-tuning the module to leverage the pre-trained model and applied it into the semantic text and visual retrieval domain. The main area of exploration for search with BERT is similarity. Similarity between documents for recommendations, and similarity between queries and documents for returning and ranking search results. Modern search engines compute the relevance of a document to a query and not just the simple overlap in words between the two queries. This is accomplished using text similarity by creating useful embeddings from the

short texts and calculating the cosine similarity between them.

### **Image Feature Representation in Computer Vision**

In the computer vision domain, the image representations derived from pre-trained Convolutional Neural Networks (CNNs) have become the new state of the art for computer vision tasks such as instance retrieval from image archives. A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics. In machine learning, deep learning networks have multiple non-linear hidden layers and can represent the data in a hierarchical way from lower to higher abstraction. CNNs are a variant of the multilayer perceptron, which are inspired by the visual cortex and have deep architectures [116,117].

#### **2.4.2 Deep Learning in Text-Image Retrieval**

This refers to the application of text-Based image retrieval applied to retrieve images, through text description as an input, which could be keyword, date, etc. Although text-Image retrieval has a wide application in semantic searching, it still poses a challenge. Text-Based image retrieval systems can be divided into two categories: single-direction retrieval and bi-directional retrieval. In single-direction retrieval, similar images to the text query are retrieved from the training set based on their similarities with the query in a trained visual space [118, 119]. In bi-directional retrieval, most approaches [120, 121] are building a common multimedia embedding space for the visual and textual data based on a training a set of image-description pairs. This matching visual data and natural language has been a challenging problem in multimedia and it has facilitated large-scale

applications, including, image and video captioning [122,123].

### **Image Captioning**

Image captioning has become an active research topic in recent years and its main contribution is that it translates multimodal data from unstructured image data to structured text data. Currently, automatic image and video captioning is extremely important, as well as challenging, because it connects the domain of computer vision and natural language processing. The current existing image captioning methods include sentence-template based [124], retrieval based [125], neural-network based [123]. The neural-network framework with the encoder-decoder structure is widely used in image caption generation, in which the encoder extracts the image features by Convolutional Neural Network (CNN), and the decoder adopts Recurrent Neural Network (RNN) to generate the image description.

### **2.4.3 Deep Learning in Human Object Interaction Extraction**

The recognition of Human-Object Interactions (HOI) [126] is a combination of visual instance recognition and the human action. Human-Object Interactions (HOI) can be represent as a triplet as ( $\langle$ Human, Verb, Object  $\rangle$ ) and the detection of this triplet is of interest in application and research. In lifelog retrieval task, the deeper temporal image understanding is still a challenge and as of yet unsolved in computer vision. Detecting a Human-Object Interactions (HOI) can assist in better understanding of the semantic meaning of image content. Currently the detection of human actions and object instance interactions with [127] has received most research focus. In addition to using object instance appearances, encoding the relative spatial relationship [128] between a person and the object with a CNN has shown improvements in detection rates. Recently, Xu [129] extends the research into HOI extraction in social image scenes and utilizes human pose and gaze estimation to improve their HOI detection model. Besides, Gao [130] introduces an instance-centric attention module which can learn to dynamically

highlight regions in an image conditioned on the appearance of each instance. Our work builds upon these recent advances in HOI detection, compared with human annotated image data, as will be described later.

## **2.5 Chapter Summary**

This chapter presents the literature review for our research work in this thesis. We investigated the conventional information retrieval methodology and lifelog retrieval system research. Especially we introduce the brief history of lifelog retrieval task and review the various approaches for solving this search problem. we detailed the lifelog disciplines and the application, analysed the existing approaches of lifelog information retrieval and highlight how our work is different with current methods. According to the literature, we know there are three important parts will affect the performance of interactive retrieval system: instances/concepts detection, semantic query-document matching method and query analysis. These three parts will be the main direction of our research.

## Chapter 3

# Research Methodology and Evaluation Methods

### 3.1 Research Methodolog

In general, research tries to find solutions and answers to questions concerning the unknown. In 2012, Creswell in [131] defines research as a logical process of steps applied to collect and analyse data in order to improve the knowledge and understanding of a topic or to solve an identified issue. Many different approaches on how to conduct research exist [131]. Based on these literature studies and the nature of the lifelog data, we decided to base our effort on the action research paradigm (summarised in Figure 3.1), as follows: We start with the problem identification, in this case, is the lacking of good retrieval system for lifeloggers to retrieve their personal moments or events. We then try to explore the research by designing how to approach the problem by conducting the literature study. Next, we collect the appropriate data (lifelog datasets in this study) and analyse them. Based on these preliminary analysis, we propose a range of possible solutions. Finally, we will try to evaluate the solutions and validate it if the performance achieve the desired outputs or repeat the whole process with the revised definition of the problem and revised solutions. Finally, after one or a number of cycles, we will either prove or disprove our hypothesis and answer the research questions.

In particular, in order to answer the four research questions mentioned in Chapter 1, we do the following experiments and the workflow of these experiments approaches

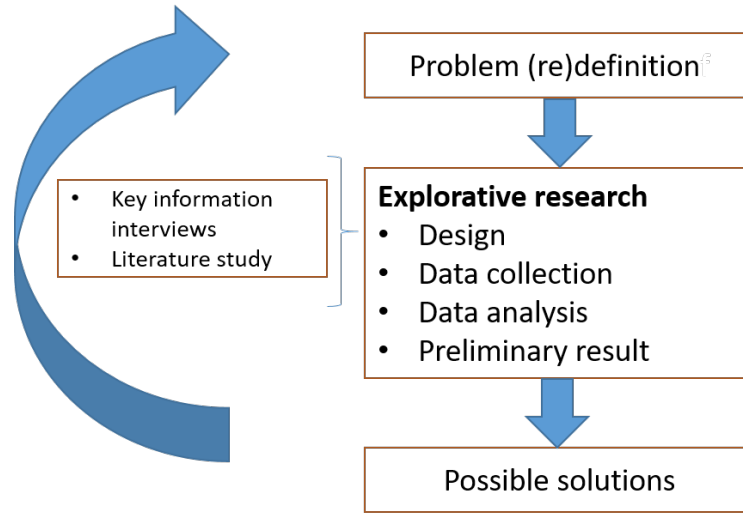


Figure 3.1: Research methodology schema.

are briefly described in Figure 3.2:

- Define what is lifelog data. Given that lifelogging is a relatively new research area, we need to define what kind of data that should be collected, and explain why and how to gather these types of data. From the collected data, propose solutions on how to organise and exploit these data in a privacy-aware and legally-compliant manner. This will create new test collections and answer RQ1 in Chapter 4.
- Build a baseline search engine, a system that allows a user to retrieve the basic moments from the lifelogs in a reliable and efficient manner, while keeping the personal life archive private and protected. This baseline search engine helps to understand how various retrieval methods (non-AI based) can be applied to this new type of data and how well they work. A detailed explanation about how to build and upgrade a baseline search engine is covered in Chapter 5.
- Improve retrieval performance by enriching the standard visual lifelog features by integrating causality-inspired features. In lifelog retrieval, there are some semantic relations, called *Causality*, that can exist between different human behaviors. Establishing such causal relations between variables from

observation of their behaviour over time is central to this scientific investigation and it is the basis for the construction of useful models and tools capable of prediction which can be used for human behavior-based retrieval. The detailed explanation about how to enhance the retrieval performance by qualifying the causality between different HOI is answered in Chapter 6.

- Integrate embedding matching techniques to enhance the retrieval performance to test whether it is beyond state-of-the-art retrieval techniques. The performance of lifelog retrieval system can be enhanced by supporting higher-level semantic mapping of a user information need into an embedding representation. The text embedding or joint embedding can be learnt to predict and match the retrieval result based on the user query. A description is presented of how to integrate text and visual-semantic embedding into lifelog retrieval in Chapter 7.

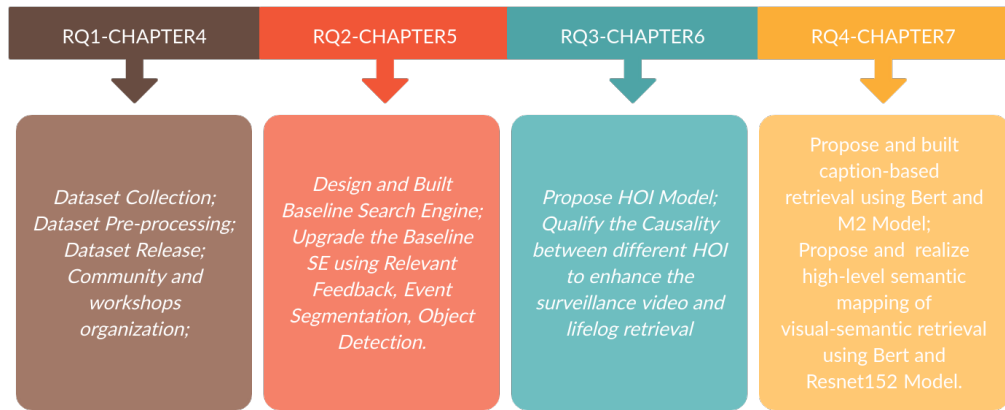


Figure 3.2: Workflow of research questions and approaches

## 3.2 Operating Constraints

As with any new topic of research, we should clearly define the constraints within which we operate. We can identify the constraints thus:

- Multimodal lifelog data should consist of a range of data that can reasonably be gathered by willing and interested individuals without any manual input required, i.e. passive capture using wearable and software sensors.
- The lifelog data collected should be made into a reusable dataset and support further lifelog retrieval research and support comparative evaluation.
- Comparative evaluation of lifelog data must respect laws and data governance laws of the jurisdiction in which the research is performed and ethical approval must be sought from the research institution.
- Any information needs for known-item search from lifelog data should be generated by the lifelogger and not subject to influence by traditional information retrieval expectations.
- Reasonable state-of-the-art data enrichment processes should be used when analysing and enriching the lifelog data, and that the focus of this research is not to develop new analytic and annotation approaches, but rather to develop new retrieval models that operate using state-of-the-art approaches.
- A reasonable state-of-the-art lifelog retrieval system should be chosen for comparison with our built systems.

These constraints are maintained for this PhD research and act as limiting factors to focus the research effort.

### 3.3 Evaluation Methods

In this section, we describe the evaluation methods employed in the research presented in this thesis. As we know, evaluating research in a correct manner is non-trivial and it is easy to prepare invalid evaluations using incorrect data or over-fit algorithms by not creating accurate datasets. Inspired by the criteria in action research proposed by Creswell in [131], we propose to consider the following criteria for the evaluations:



- Impact and Usefulness. Does the research impact to the community (in this case, lifelogger)? Does it help in solving a problem, empowering them, changing them, or providing new understandings?
- Data. How good is the data collection process? Does the collected data help to address the problem?
- Achievements. Reports the research in a way that is understandable and useful to audiences. In particular, what are the targets (conferences and journals) of the research?
- Collaboration. Collaborates with others during the study to find the best solutions and compare my approaches to the existing state-of-the-art.
- Novelty. Does the research advance a plan of action for trying to solve the problem?

We took these criteria into account when developing known-item search engines for lifelog data in an open comparative process where we could test our systems against other researchers in an open and comparative process. We achieved this by developing the test collections (datasets with queries and relevance judgements) and releasing them through agreed data governance and ethical approved processes. These test collections were used by participants in three of the main comparative benchmarking exercises (NTCIR, ImageCLEF, LSC). In terms of evaluation, we use standard retrieval evaluation measures, which we outline in next subsection.

### 3.3.1 Traditional Evaluation Methods

In our research, we evaluate our approach using Trec\_eval software which is available at: ([https://trec.nist.gov/trec\\_eval](https://trec.nist.gov/trec_eval)). It uses different measures of information retrieval. In our work, we choose to use F-score and MAP as evaluation metrics. The relevant evaluation approaches are described below.

## Precision and Recall

Precision and Recall are standard approaches to evaluation in Information Retrieval(binary-rank-order), Machine Learning and Pattern Recognition. Precision, which can also be called positively predicted rate, is the fraction of corrected retrieved instances 3.1, while Recall (also known as Sensitivity, see Equation 3.1 ) is the fraction of relevant instances that are retrieved. Therefore, both precision and recall are based on understanding and measure of relevance of programs. Generally speaking, precision is the probability of how relevant a retrieved item is, while recall is the probability of whether a relevant item is retrieved in a search. The relation between precision and recall is shown in Figure 3.1.

Table 3.1: Precision\_Recall

	Relevant	Non-relevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

$$precision = \frac{TP}{TP + FP}; recall = \frac{TP}{TP + FN}; \quad (3.1)$$

## F-score

In order to determine if our tests are significant, we conduct some test statistics. In our case, we use the F statistic, also called F-measure. F value is a single measure that represents the harmonic mean of precision and recall, see Equation 3.2. In this case, recall and precision are evenly weighted, so it is also called F1 value. It conveys the significance of tests or evaluations.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (3.2)$$

## AP and MAP

Precision and recall are single-value metrics to measure the performance of information retrieval systems. They both measure the holistic list of documents returned by a program or system without considering the ranking order of the returned documents. Average precision, however, is a measure of the average value of  $p(r)$  in precision-recall curve over the interval from  $r = 0$  to  $r = 1$ , which means it puts the ranking order of retrieved items into consideration. If we generate a list of precision-recall values based on different chosen  $N$  and draw a curve to represent every precision-recall dots for top  $N$  retrieved items, this curve is called precision-recall curve. This curve plots precision  $p(r)$  as a function of recall  $r$ . Average precision is the average value of all precision values. In Equation 3.3,  $P(k)$  is the precision of the  $I_k$  in the retrieved item list, and  $relk$  is 1 if  $I_k$  is relevant, otherwise, it is 0.

Mean Average Precision (MAP), however, is the most commonly used measurement for retrieval tasks among the TREC community. It provides a single-figure measure of system quality across all precision levels, see Equation 3.5. Among all evaluation measures, MAP has been proved to be very in anti-discriminating and stable.

$$AP = \frac{1}{|I_{rel}|} \cdot \sum_{k=1}^n P(k) * rel(k) \quad (3.3)$$

$$rel(k) = \begin{cases} 1 & \text{if } I_k \text{ is relevant} \\ 0 & \text{if } I_k \text{ is not relevant} \end{cases} \quad (3.4)$$

$$MAP(Q) = \frac{1}{Q} \cdot \sum_q^Q AP(q) \quad (3.5)$$

### 3.3.2 Enhanced Evaluation Methods

In order to assess performance, we deployed classic metrics, including:

- Cluster Recall at X (CR@X) - 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.

We have used various cut-off points including  $X=5, 10, 20, 30, 40, 50$ . The main ranking metric is the F1-measure@10, which gives equal importance to diversity (via CR@10) and relevance (via P@10).

## Chapter 4

# Datasets Collecting and Preprocessing

### 4.1 Introduction

In this chapter we address RQ1, which is **how to construct a test collection of lifelog data to support retrieval system comparison and how to create this test collection in a privacy-aware manner?**

In order to answer this question, we define what is lifelog data and construct a lifelog dataset based on the process which is detailed in this chapter. Further, a list of evaluation methods adapted for our experiments are discussed and explained.

We needed to use the test collection methodology but there were no test collections, so we had to address the challenge of how to build a disclosed lifelog dataset by proposing the principles for building and sharing such types of data. In Section 4.2, we state the research methodology employed in our research procedure and propose principles for construction of a disclosed lifelog dataset as well as describe the process of how we built a new lifelog dataset. And in Section 3.3, we describe the evaluation methods employed in this research.

### 4.2 Test Collection Construction Methodology

In order to get insights from the lifelogs, individuals or researchers require tools for knowledge extraction, search, summarising, and visualisation. Moreover, to validate

the performance of these tools, they need data, including large and rich collections of lifelog data. In this part, we will describe the detailed processes for how we built the benchmarking lifelog dataset for lifelog retrieval tasks in NTCIR, ImageCLEF, LSC lifelog. We will mainly focus on NTCIR13-lifelog dataset, since this is the main dataset used in our experiments.

#### 4.2.1 Principle of Building a Lifelog Dataset

Learning from the NTCIR-12 [33] and the LTA-2016 [132], we propose these principles for building a disclosed lifelog dataset:

- **The Continuity.** *Lifelog data of each individual should be captured continuously for at least 30 days.*

Getting insights from lifelogs could be a retrieval problem, e.g., “find the moments that a logger having dinner”, or could be also an analytics problem, e.g., “compare the eating habits between two loggers”. In order to answer these queries, the dataset must contains sufficient information, i.e., the lifelogs of each logger should be gathered continuously over a number of days. In this study, we propose that the minimum amount should be 30 days, which we learned from NTCIR-12, that a 30-day period is sufficient for identifying habit patterns of an individual.

- **The Completeness.** *lifelogs should contain four basic types of information: visual data, personal biometrics, human activity, and information accesses.*

In order to allow for statistically significant studies in the field of lifelogging, a dataset needs to be large enough to represent real-world data of lifeloggers. As addressed in the challenges, we should log all information from all sources, however, it is not doable in practice, thus, we propose these four sources as that main data that a lifelog dataset should contain: **(i) Visual information:** use for the continuous and automatic capture of life activities as a visual sequence of digital images. **(ii) Personal biometrics:** the

lifelogs should contain basic personal biometrics of the individual, e.g., heart-rate, calorie, and others. **(iii) Human activity:** there are many personal sensing devices for monitoring everyday activities, which can monitor human performance, for example activity levels (number of steps, distance traveled), sleep duration, etc. **(iv) Information access:** This could refer to using a computing device (e.g. smartphone or computer) to continuously and passively capture the user's context or logging the consumed data from all the words typed, web pages read, YouTube videos watched and so on. It is worth noticing that the NTCIR-12 dataset does not have personal biometrics and information access, and thus it does not fully satisfy this principal but subsequent datasets did, such as the NTCIR13 dataset we used.

- **The Anonymity.** *All user-identifiable data have to be removed.*

We must consider the principles of privacy-by-design [133] when creating the lifelog dataset. In order to remove user-identifiable information while maintaining the usefulness of the data, we propose these information have to be anonymised: human faces, personal IDs, vehicle plates, name tags, and the address/location of the individual's home and work place.

- **The Protectiveness.**

*The dataset should be password protected and all accesses should be logged.* the lifeloggers who donated the data, or any identifiable individuals who appear in the data can request some content to be deleted at any time and need to agree to use of data beyond any initial planned dataset release. The researchers who manage the dataset can also request data deletion at any time.

The lifelog data is anonymised, however, it contains experiences of loggers as well as individuals appeared in the lifelogs, and thus it must be only shared for research or education purposes. These information, hence, should be also protected strictly by agreement between loggers, users and the people who manage the dataset.

These principles answer the last five challenges pointed in Chapter 1.1.2 while we ourselves and some other volunteers provide the solution for the first challenge. Applying these principles, we collected the lifelog data which gathered continuously for over 30 days from rich multi-modal sensors and devices, and built new lifelog datasets: the NTCIR-12 lifelog, NTCIR-13 - Lifelog, NTCIR-14 lifelog. In the next subsection, we describe the process for building this dataset. It is worth noticing that following these principles, these two dataset are reusable, which means they can support a number of years of ongoing research activities.

#### 4.2.2 Process of Building Personal Life Datasets

In order to build a lifelog dataset that respects the proposed principles, we apply a 3-step process, summarized in Figure 4.2, as follows: firstly we gather individuals' lifelogs by following the first two principles. Next, the gathered data is cleaned by removing user-identifiable data (applying the anonymity principle) and then organized hierarchically with the basic units composed from every minute. Finally, before publishing the dataset, we protect the data by making agreements between loggers, users and the people who manage the dataset, as well as putting in place a data download tracking mechanism. These steps are described as follows:

##### Step1: Data Gathering

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 their mobile phones with two wearable devices: a Narrative Clip<sup>21</sup> and a smart-watch<sup>2</sup>see Figure 4.1). It is worth noticing that in the NTCIR-13 Lifelog dataset, each individual gathered continuously over 45 days, that fulfilled our first principle.

The data was gathered as follows:

---

<sup>1</sup><http://getnarrative.com/>

<sup>2</sup>The lifeloggers used the BASIS smartwatch.





Figure 4.1: A lifelogger with his everyday wearable devices: A Narrative 2 camera (wearing on his neck) and a smartwatch (the black watch on his right arm).

- Visual information: the visual data was gathered using a Narrative Clip 2, taking photo (landscape whenever possible) at each 45 seconds, from breakfast to sleep. This produces about 1,500 images per day from the viewpoint of wearer.
- Personal biometrics: We use the smartwatch to obtain the bio-metrics information over  $24 \times 7$ , gathering these information every second: heart rate, galvanic skin response, calorie burn, and steps. Beside such biometric information, some lifeloggers also provide their blood pressure and blood sugar, measured every morning.
- Human activity: In order to log the individuals physical activities, the Moves app is used. Moves app is a smartphone app that automatically records user activity in terms of semantic locations and physical activities (e.g., waking, 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.
- Information access: Collected using the LoggerMan app [134]. LoggerMan

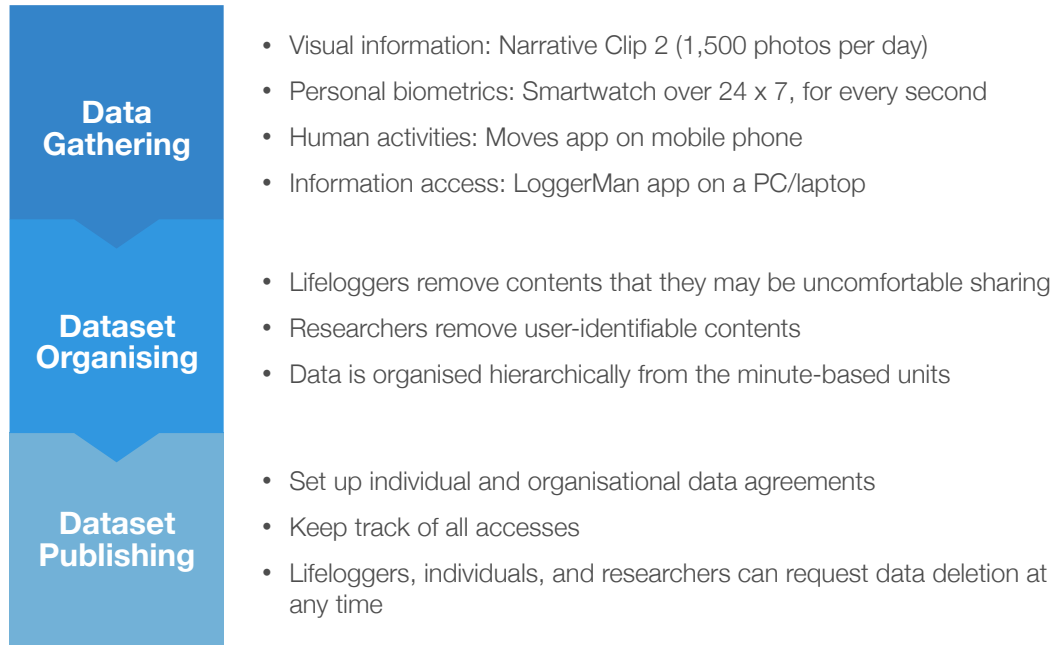


Figure 4.2: The proposed process for building a disclosed lifelog dataset.

helps researchers and lifeloggers to collect interaction data produced during normal computer usage. The main goal of LoggerMan is to work passively in the background, intercept usage events and store them for later analysis. It gathers wide range of keyboard, mouse and UI actions, thereby capturing the information creation and consumption activities of the individual

## Step2: Dataset Organising

From the gathered the data, the second step is applied to first remove user-identifiable content, i.e., apply the anonymisation process (details of this process is described in the section 4.2.3). The anonymised data is then checked to ensure temporal alignment of the sensor data, since it is gathered from different devices. It was necessary to check and resolve alignment problems (typically in the order of 1-2 minutes) for a lifelogger by cross-referencing reported timestamps from the Narrative 2 camera with clocks captured daily in the real-world. Finally, the synchronized data is organised hierarchically from the minute-based units. Typically in information retrieval (IR), there is a single basic unit for indexing and

retrieval. For many IR tasks, this basic unit is the preferred as unit of retrieval and choosing the basic unit may not trivial. In web search, a document could be considered as a basic unit, however, for other IR problems, a basic unit as a document is too large to be of value to the user. Consider video search as an example of an IR application, a video can be seen as too large to be useful, and thus, the video is often segmented into sub-units called video shots or scenes which are shorter and more likely to answers a user's information need. Back to the lifelog data, it is not trivial to decide what the basic unit is since the lifelogs are composed of different types of data, captured at different frequencies (1 second to potentially 1 day). In order to deal with this problem, we propose to organize the data hierarchically where the top levels are the days, sorted by chronological order, and each day contains 1,440 ( $24\text{ hours} \times 60\text{ minutes}$ ) minute-based units. All data is accompanied by XML markup at the minute level of granularity. An example of the XML description is shown in Figure 4.3. Verified by the NTCIR-12 task, it is confirmed that this structure is useful for analytics approaches to get insights from the lifelog data.

### Step3: Dataset Publishing

For dataset publishing, we set up agreements between lifeloggers, users, and researchers/people who manage the dataset and keep track for all accesses. We also define four components (summarized in Figure 4.4) of the published dataset as follows:

- Researcher/People who manage the dataset: do the anonymisation and organise the data.
- Storage: Beside hosting the dataset, this “component” also protects the data by applying password protection for each data file. We also protect the data by making individual and organisational data agreements.
- Dataset: the lifeloggers and individuals can request data deletion at any time

```

<minute id="625">
  <location>
    <latitude>53.360523</latitude>
    <longitude>-6.2604883</longitude>
  </location>
  <activity>transport</activity>
  <bodymetrics>
    <calories>1.5</calories>
    <gsr>6.65E-05</gsr>
    <heart-rate>82</heart-rate>
    <skin-temp>88.7</skin-temp>
    <steps>0</steps>
  </bodymetrics>
  <images>
    <image>
      <image-id>u1_2016-08-08_092502_1</image-id>
      <image-path>u1/2016-08-08/20160808_092502_000.jpg</image-path>
    </image>
    <image>
      <image-id>u1_2016-08-08_092549_2</image-id>
      <image-path>u1/2016-08-08/20160808_092549_000.jpg</image-path>
    </image>
  </images>
</minute>

```

Figure 4.3: An example of the XML data description for one minute

and need to agree to use of data beyond initial plan. The researchers can also request data deletion at any time.

### 4.2.3 Data Protection and Privacy

How to support sharing and privacy-aware analytics is a major issue in lifelogging, and thus, we use this subsection to discuss this issue as well as describe the anonymisation process we designed and applied in building all lifelog dataset.

#### Data Level

Lifeloggers who donate their dataset are willing to share their everyday activities, however, we felt it was necessary to give them an opportunity to remove any data that they may be uncomfortable sharing. This required a manual inspection of all their lifelog data before sharing it with the another researchers. After this, all images

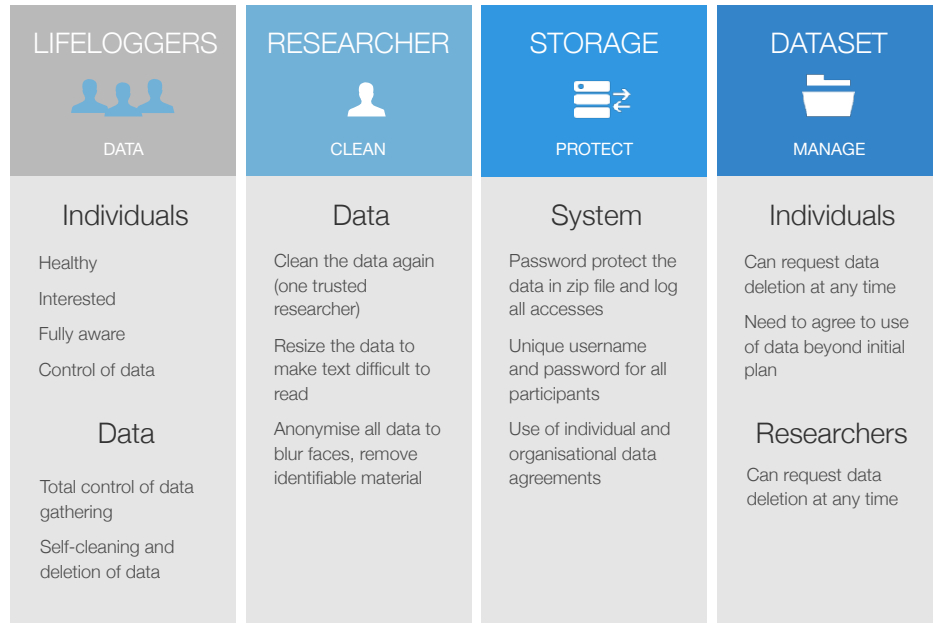


Figure 4.4: The components of a published lifelog dataset.

were reviewed by a trusted researcher with oversight of the entire dataset to ensure that no potentially embarrassing or offensive images were accidentally included in the dataset.

To ensure privacy of both the lifeloggers and individuals (subjects and bystanders) captured in the lifelog, user-identifiable content is removed by the following process (summarized in Figure 4.5 where some visual anonymisation results can be seen in Figure 4.6): First, each recognisable face in every image was blurred in a manual process, which ensured no false positives or missed faces. It is worth noticing that this process has to be done manually since the available face detection methods were not capable of detecting every human face in the photos. Car plates, handwritten notes, or any other user-identifiable contents were also manually blurred from the images. Next, the images were resized down to  $1024 \times 768$  resolution, in order to reduce the storage space as well as making the majority of any on-screen text captured by the lifelogging camera unreadable.

Similar rules are applied on the information access and creation data collected by LoggerMan [134], all user-identifiable contents are removed from the information

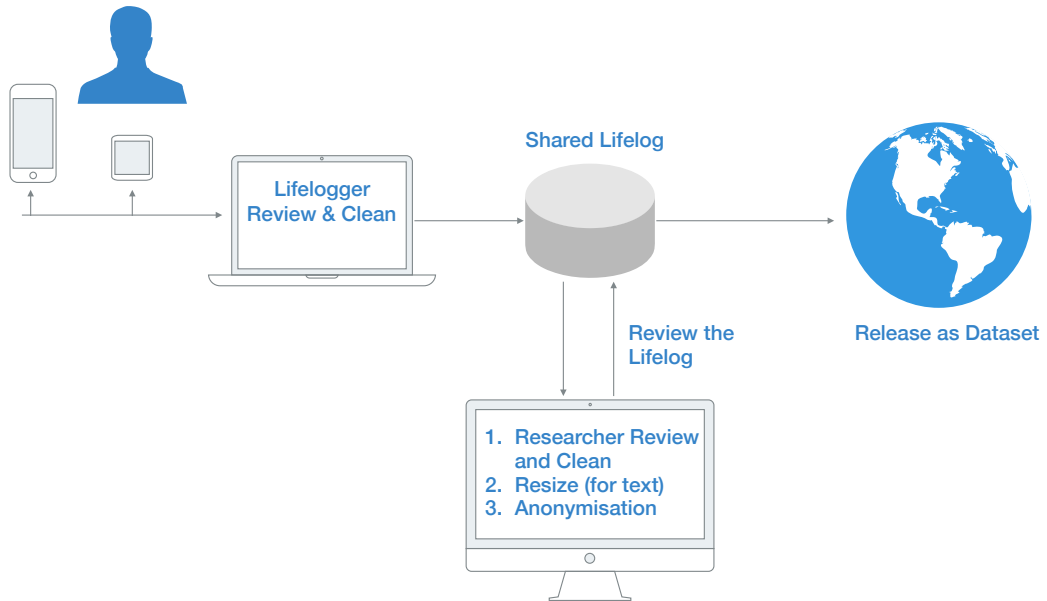


Figure 4.5: The anonymisation process applied on raw lifelog data.

access data.

Since privacy extends beyond content of 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.

## Application Level

Beside the anonymisation at data level, we also propose a model for protecting the privacy at the application level, i.e., allow doing analytics without working on the raw data. If needed this is done by a classic way (summarized in Figure 4.7): providing meaningful features instead of the raw data. Among different types of features, we decided to exploit the Microsoft Computer Vision API<sup>3</sup> to extract the information about visual content found in the lifelog images. This API allows to identify content and label it with confidence, which allows different approaches can get insights from the lifelogs without needing the raw data. These semantic features are published

<sup>3</sup><https://www.microsoft.com/cognitive-services/en-us/computer-vision-api>



(a) A human face and his name tag.



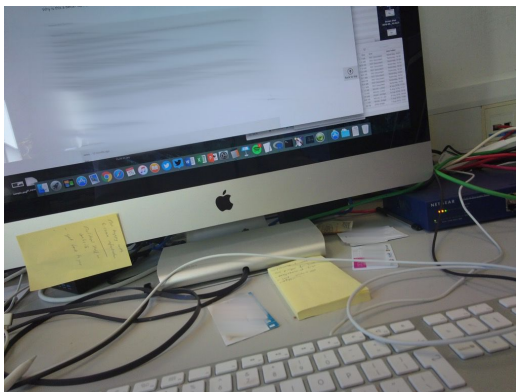
(b) A car plate.



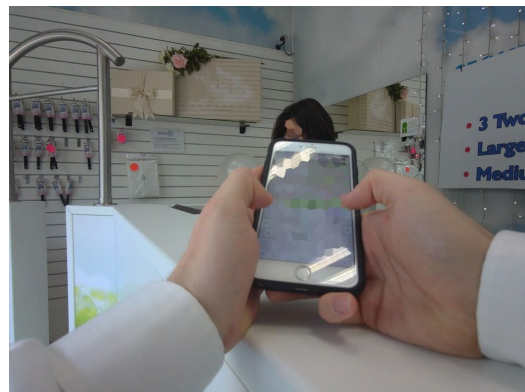
(c) Handwritten notes.



(d) Human faces.



(e) Monitor.



(f) Contents on mobile screen.

Figure 4.6: Examples of the visual anonymisation.



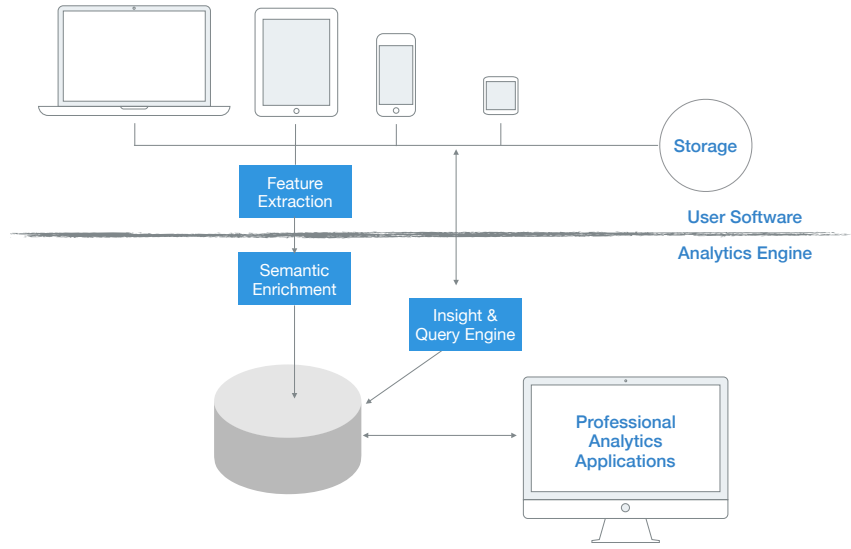


Figure 4.7: A proposed privacy-aware lifelogging model.

together with the images in the NTCIR-13 - Lifelog dataset.

#### 4.2.3.1 Benchmarking Initiatives and Workshops on Life Archive Analytics

Good IR systems need comparison, so we co-organised the NTCIR tasks, so that baseline can be compared. Researchers usually need data to evaluate their methods, and there is no exception for researchers in personal life archive organisation and retrieval. To support such research efforts, large volumes of lifelog data have been gathered from several volunteer lifeloggers and organized, annotated and published them for researchers as the lifelogging tasks in NTCIR 12 - Lifelog, NTCIR 13 - Lifelog 2, NTCIR 14 - Lifelog <sup>24</sup> and Imageclef-17, Imageclef-18, Imageclef-19, Imageclef-20<sup>5</sup> and LSC-18, LSC-19, LSC20<sup>6</sup>. To the best of our knowledge, these collections are the largest (in terms of number of days and the size of the collection) and richest (in terms of types of information) datasets on personal life archives ever shared. These datasets used in are referred to the task overview papers for NTCIR-

<sup>4</sup><http://ntcir-lifelog.computing.dcu.ie/>

<sup>5</sup><https://www.imageclef.org/2020/lifelog/>

<sup>6</sup><http://lsc.dcu.ie/>



Table 4.1: Statistics of NTCIR Lifelog Datasets

Criteria	NTCIR-12	NTCIR-13	NTCIR-14
Number of Lifeloggers	3	2	2
Number of Days	90 days	90 days	43 days
Collection Size	18 GB	26 GB	14 GB
Number of Images	88,124 images	114,547 images	81,474 images
Number of Locations	130 locations	138 locations	61 locations
Physical Activities	Moves app	Moves app	Moves app
Calorie Burn	-	Fitness Watch	Fitness Watch
Step Count	-	Fitness Watch	Fitness Watch
Heart Rate	-	Chest Strap	Fitness Watch
Blood Glucose	-	Daily	Continuous
Music Listening	-	Last.FM	Last.FM
Cholesterol	-	Weekly	-
Uric Acid	-	Weekly	-
Diet Log	-	Manual	Manual
Conventional Photos	-	-	Smartphone

12 [40], NTCIR-13 [33] and NTCIR-14 [34]. See Table 4.1 for a summary comparison of the three datasets. Event though we built and released these three datasets, but we will only focus on NTCIR-13 lifelog since it have richest data and we used it for our research experiments.

In order to better view the relationship between tasks and dataset, we summarised our evaluation benchmarks and the dataset used in Table 4.2. The work we maninly focused on is LSAT in NTCIR-lifelog and LMRT in Imageclef-lifelog retrieval task. According to the table, we can confirm the proposed dataset is enough for the proposed topics, in which the answers can be achieved by exploiting the provided multimodal data.

Over the last 4 years, we identified the challenges for building a shared personal life archive dataset [135], proposed principles, built and described the whole processes from data gathering to determining the roles for the people who are building, sharing and exploiting such kind of data [25]. These principles can be considered as references for systems that collect personal life archives.

As the first step of building the community working on personal life archives, we

Table 4.2: Statistic of Lifelog Comptition and Tasks

Criteria	Competition	Tasks	Topics	Description
NTCIR12	NTCIR12 Lifelog	LRT	48	Know Item Search Task
		LIT	10	Gain insights into lifelogger's daily life
	ImageClef2017 Lifelog	LRT	16	Know Item Search Task
		LST	5	Analysis and Summarize data based on requirement
NTCIR13	NTCIR13 Lifelog	LSAT	24	Explore search and retrieval from lifelogs
		LAT	15	Explore computer vision algorithms to describe the lifelog images.
		LEST	15	Explore knowledge mining and visualisation of lifelogs
		LIT	5	Gain insights into lifelogger's daily life
		LMRT	20	Retrieve a number of specific moments in a lifeloggers life
		ADLT	20	Analyse the lifelog data and provide a summarisation based on concept
NTCIR14	ImageClef2018 Lifelog	LSAT	24	Explore search and retrieval from lifelogs
		LADT	16	Identify Activities of Daily Living from lifelogs,
		LIT	N/A	Explore knowledge mining and visualisation of lifelogs
	NTCIR14 Lifelog Task	LMRT	20	Retrieve a number of specific moments in a lifeloggers life
		Puzzle	20	Analyse and predict images without timestamps

increasingly organize related workshops and panels: iConf 2016<sup>7</sup>, Lifelogging Tools and Applications in ACM MM 2016 [132] and ACM MM 2017<sup>8</sup>. Together with these, we organize rigorous comparative benchmarking initiatives: NTCIR 12 - Lifelog [32, 40], NTCIR 13 - Lifelog 2, NTCIR 14 - Lifelog, LifeLog task at ImageCLEF 2017 [136], ImageCLEF 2018 [137], ImageCLEF 2019 [138] and the ImageCLEF2020 [138], LSC18 [139], LSC19 and LSC20 [140] which aim to bring the attention of personal live archive analytics to a wide audience and to motivate research into some of the key challenges of the field.

---

**Listing 1** Description of selected topics in Imageclef2019 LMRT subtask.

---



---

**T.001 Icecream by the Sea**

---

Description: Find the moment when u1 was eating an icecream beside the sea

Narrative: To be relevant, the moment must show both the ice cream with cone in the hand of u1 as well as the sea clearly visible. Any moments by the sea, or eating an ice cream which do not occur together are not considered to be relevant.

**T.002 Having Food in a Restaurant**

---

Description: Find the moment when u1 was eating food or drinkingU1 was eating food in a restaurant while away from home. Any kinds of dishes are relevant. Only Drinking coffee and have dessert in a cafe won't be relevant.

**T.003 Watching Videos**

---

Description: Find the moment when u1 was watching video when using other digital devices.

Narrative: To be relevant, u1 must be watching videos in any location and any digital devices can be considered. For example: TV machine, tablet, mobile phone, laptop, desktop computer.

**T.004 Photograph of a Bridge**

---

Description: Find the moment when u1 was taking a photo of a bridge.

Narrative: U1 was walking on a pedestrian street and stopped to take a photo of a bridge. Moments when u1 was walking on a street without stopping to take a photo of a bridge are not relevant. Any other moment showing a bridge when a photo was not being taken are also not considered to be relevant.

**T.005 Paper or Book Reviewing**

---

Description: Find all moments when u1 was reading a paper or book.

Narrative: To be relevant, the paper or book must be visible in front of U1 and sometimes U1 use a pen to mark on the paper or book.

---

Typically, for each benchmarking initiative, together with the dataset, we introduced several tasks which aims to advance the state-of-the-art research in lifelogging as an application of information retrieval. For example, in ImageCLEFlifelog 2017 edition [58], we introduce two tasks: Lifelog Retrieval Task

---

<sup>7</sup><http://irlld2016.computing.dcu.ie/index.html>

<sup>8</sup><http://lta2017.computing.dcu.ie>

**Listing 2** Description of selected topics in NTCIR13-lifelog LSAT subtask.**T.001 Eating Lunch**

Description: Find the moments when I was eating lunch.

Narrative: Moments in which the user was eating lunch are relevant regardless of where the lunch is eaten.

**T.002 Gardening**

Description: Find moments when I was gardening in my home.

Narrative: Relevant moments should show the user in the garden and interacting with vegetation in some way, such as trimming bushes or cutting grass. The gardening activity must take place in the user's home and not any other location.

**T.003 Castle at Night**

Description: Find the moment when I was visiting a castle at night.

Narrative: To be relevant, the moment must show a castle clearly and it must be at night.

**T.004 Coffee**

Description: Moments that show the user consuming coffee or tea in a cafe (outside of home or office) are considered relevant. The coffee can be hot in a cup or paper cup, or cold coffee in a plastic or paper cup.

**T.005 Sunset**

Description: Find the moments when I was outside at sunset.

Narrative: To be considered relevant, the moment must show the sun setting. This can be at night time, or can be when the sun is disappearing behind a mountain in the evening. Sunrise moments are not considered relevant.

(LRT) and Lifelog Summarisation Task (LST). In LRT, the participants had to analyse the lifelog data and for several specific queries, return the correct answers, for example *"Find the moment(s) when user u1 was having a drink in a bar with someone". To be relevant, the user must clearly be inside a bar. Any moment in which the user is clearly seen having a beer or other drink in a bar venue is considered relevant. Having a drink in any other location(e.g. a cafe), or without another person present is not considered relevant. The type of drink is not relevant once it is presumed alcoholic in nature and not tea/coffee. Blurred or out of focus images are not relevant. Images that are covered (mostly by the lifelogger's arm) are not relevant.*

Listing1 and Listing2 show the selected retrieval topics proposed in NTCIR13-Lifelog, ImageCLEF 2019 .

### 4.3 Discussion & Contribution

We reported challenges and proposed principles for building a disclosed lifelog dataset in this chapter. This is the first time the problems of building such a dataset has been discussed in detail.

It is an inevitable fact that the main challenge for benchmarking organization is the effort required to generate appropriate and real-world datasets and release them in an ethically and legally complaint manner. The three lifelog datasets released at NTCIR represent about a year of effort in total from a number of researchers and lifeloggers; this naturally incurs significant expenses in terms of organisers time and resources. Real-world use-cases are likely to either focus on retrieval from longitudinal archives donated by one individual, or across large populations (as in epidemiological studies) and the data gathering and release methodology employed for this task was not ideal, due to the large overhead of effort required to ensure privacy-preservation.

In this work, our main contribution is the identification of the challenges for building a shared lifelog dataset and the proposition of principles for building a shared lifelog dataset. Meanwhile, based on the proposed principles, we built and described the whole process from data gathering to determine the roles for the people who are building, sharing and exploiting a disclosed lifelog dataset, and released the three disclosed datasets for widespread use. At last we proposed how future lifelog data should be stored and organised, and therefore made easy for a user to access via an appropriate search mechanism.

### 4.4 Chapter Summary

In this chapter, we address the research question of how to construct a test collection of lifelog data to support retrieval system comparison and how to create this test collection in a privacy-aware and legally-compliant manner. In order to answer these question, we present the research methodologies to help us achieve the research goals

during this work. We presented a set of possible challenges and opportunities from personal life archives. We proposed the principles for building and sharing such types of data, which is the first time the problems of building such kind of dataset has been deeply discussed. We identified methods and technologies that can aid users to get insights in their data, from public resources to their social connections and closer intimacy, and moreover, from their personal life archive, including different granularities of information from their past experiences. We presented the current state of the field and provided a number of potential benefits. Furthermore, we identified and proposed solutions to the challenges that arise with such data. This will be increasingly important over the coming years as we learn more about ourselves and have access to technologies that will help us in many aspects of everyday life. We also showed that there are several benchmarking and works in lifelog archive analytics we organised. Finally we presented the traditional IR evaluation method and enhanced evaluation methods we adopted in our experiment.

In the next three chapters, we present our research work on improving know-item search for lifelog collections and show the experiment results to demonstrate the effectiveness of our proposed work.

## Chapter 5

# Baseline Search Engine for Lifelogs

### 5.1 Introduction

In this chapter we address RQ2, which is **how to apply traditional information retrieval techniques on this multi-modal lifelog data to develop a high-quality baseline search engine?**

To answer this research question, we have proposed and developed the baseline search engine for lifelog data collection, which aims to make the lifelog data searchable, organisable and easy to be updated and served as baseline comparison for other lifelog retrieval systems that took part in the various bechmarking activities we organised.

In real life, the lifelog data generated by lifelogging tools and software is usually noisy, has errors, is unaligned, unorganised, and essentially overwhelming for the individual [26]. This is due to several factors, some of them are: 1) the variety of devices and sensors used for data gathering, which commonly are incompatible and from different companies/manufacturers. 2) the long and intense nature of the data logging process, that can easily result in missing data and/or faulty sensors. 3) the huge amount of data in different formats, representation and time stamps. This raises the need for more research and effort to be invested into the development of lifelog retrieval systems that can address these challenges and meet the potential opportunities such comprehensive personal data promises. One of the interesting use cases is data retrieval and accessibility applications: allowing the individual to search and access previous life events in efficient and timely manner. This is a yet

unsolved challenge and to address this, collaborative benchmarking exercises and workshops have been organised by us recently such as NTCIR-13 - Lifelog [33] and ImageCLEF2017lifelog [141]. We noted that there was still no standard or typical approach to developing a lifelog search engine arising from these benchmarking exercises.

In this chapter, we motivate the need for, and present a baseline search engine for personal life archives, which aims to make the personal life archive searchable, organisable 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. In order to provide an efficient retrieval system in lifelog retrieval domain, based primarily on faceted querying using captured metadata, which served as a baseline for other systems, we developed a prototype interactive lifelog search engine for use in answering interactive search topics. The main idea behind this baseline search engine is to provide a starting point for researchers in the area and provide a baseline for comparative evolution.

In order to implement an action research cycle to produce a good baseline search engine by activity, we took part in many benchmarking efforts including NTCIR13, Imageclef 2017, Imageclef2018, Imageclef2019 lifelog tasks. Then we gained experience and knowledge from other team's work to improve and develop a better system. The cycle pathway for system design and enhancement are shown in Figure 5.1 and the detailed description about these systems are in the following subsections. These 4 systems all employ the minute as the basic unit, but experimental approaches in system 2 are also designed using image-based and segment-based unit for comparison. Significantly, the first 3 systems were fully designed and developed from scratch by the dissertation author, and LIFER2 was designed by the author and developed by a wider collection of our team members.



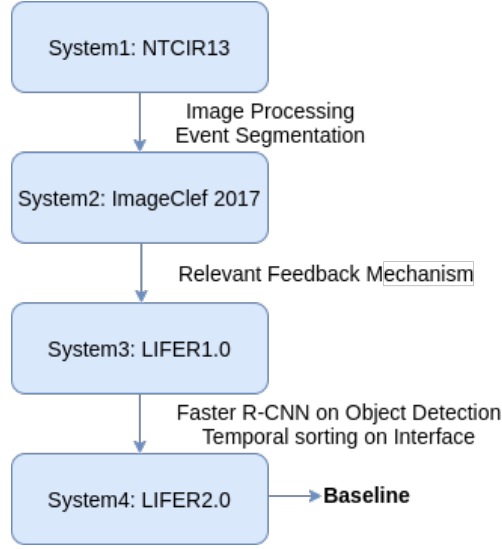


Figure 5.1: The learning improvement and updating steps of the lifelog baseline system

## 5.2 System1: NTCIR13-lifelog

The main idea behind this baseline search engine was first to provide a starting point for researchers in the area, as well as a documented system for comparative analysis. The baseline search engine extracted textual data from the raw data for higher level analysis. Then, it indexed all the extracted features and hierarchically organised them which allow queries to be processed via an API/interface.

### 5.2.1 Search Engine Architecture

Figure 5.2 summarises how we designed the baseline search engine system [142] based on best practice from the provided raw data, we extract locations, visual concepts, time, and activities and transformed this raw data into index-able feature vectors. These feature vectors are then indexed and hierarchically organised. Finally, a user or other system can define a faceted query and retrieve ranked moments via the interface.

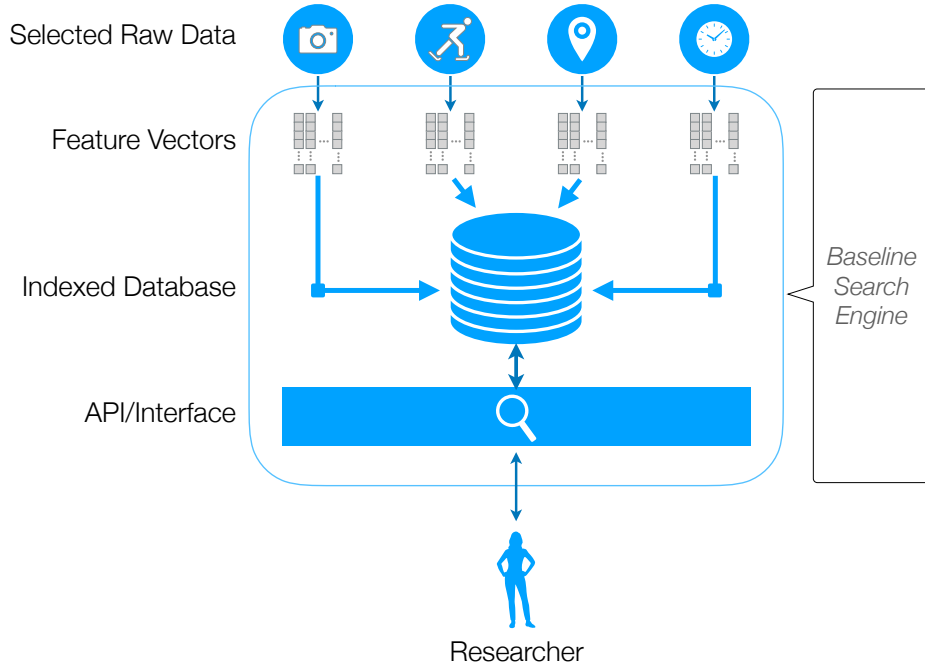


Figure 5.2: The baseline search engine architecture

### 5.2.2 Data Organisation and Retrieval Process

To organise and index the data, we follow the study in [142] by arranging features as chronological order, and use the minute as the basic unit. Building up from these basic units, we organise the data at higher level which can be turned into more useful information, the minutes are hierarchically grouped into event nodes (typically, in a full day, a person encounters anything upwards of 20 individual events, with each lasting (on average) 30 minutes [43,44]), then ultimately leading to larger units such as days and multi-day events (e.g. holidays).

The retrieval is then simply done by ranking moments (in this task, images) that matched with the queried criteria. The challenge here is how to turn a query into specific criteria. It can be automatically done by (1) considering any word in the topic as the queried concepts and then searching for all images that contain those concepts (2) fine-tuning the query manually by “a human (in the loop)”, i.e., the researcher (we) will read the topic and “translate” it into the search criteria. For this task, we applied both methods and the results will be reported in the section 5.2.4.

For example, with the query: Some examples can be seen in Figure 5.3 and the query is translated as below:

“Find the moments when user u1 was using public transportation or taxi at sunset.”

We follow the study methodology proposed in [141] and “translate” that query into specific required pieces of information, as follows:

- User = {u1},
- Concepts = {sunset},
- Activity = {transport, airplane},
- time = {16:00-21:00}
- Location = {n/A}



Figure 5.3: Examples of the results retrieved by the proposed baseline search engine for the query “Find the moment when I was taking public transportation or taxi at sunset.”

### 5.2.3 Ranking

To refine the results, i.e., to increase the precision of the top retrieved images, we use a hierarchical agglomerative clustering algorithm (see [143]) to group similar images into the same cluster based on all of their features. The clusters are then sorted based on the number of images, in decreasing rank order. Finally, we produce the retrieval by selecting representative images from the clusters by choosing the images closest to the center of each cluster.

### 5.2.4 Results

We applied the baseline search engine to the LSAT sub-task in two submissions of NTCIR13-lifelog competition: the first one is a fully automatic approach by taking all words in the query as the searching criteria. In the second submission, we manually “translate” the topic into criteria. With automatic method, the official result is obtained at 0.098. With the human-in-the-loop method in the second submission, we got a score of 0.329. Looking into details, the precision and recall at cut-off point at 10 we got are 0.48 and 0.11 by using human-in-the-loop method, respectively. For some topics, the results reaching the “golden results”, i.e., the score at the ground-truth, for example at topic LSAT003 (“Find the moment when I was visiting a castle at night”) we got 1.00 and 0.92 for precision and recall, respectively, while at some topics, for example LSAT019 (“Find the moment when I was painting the walls at home”), the results were all zeros at cut-off point at 10. It shown that further searches are required to improve the retrieval on this kind of data.

### 5.2.5 Learning Improvement

In this section, we introduced methods for retrieving basic moments from the lifelog data in a reliable and efficient manner by exploiting the baseline search engine. At the end, we planed to enrich, and extend the search engine with more agile and advanced solutions, aiming at giving better information for higher level of insight

and query engines. Firstly, the lifelog images which are blurred or covered can be filtered and removed. Then, in order to better track the fine-grained events in lifelog data, we choose to use chronological-based segmentation to identify the temporal event boundaries. These are the learning improvements for the system2.

### 5.3 System2: ImageClef2017

This section described our participation in the ImageCLEFlifelog 2017 Retrieval subtasks. In this section, we proposed some baseline approaches, using only the provided information, which require different involvement levels from the users. With these baselines we target at providing references for other approaches that aim to solve the problems of lifelog retrieval.

#### 5.3.1 Proposed Approaches

The proposed approaches follow the schema as illustrated in Figure 5.4. Since lifelogs are chronologically organised and moments in the same activity or the same event are normally very similar, we group similar moments together based on time and concepts. By applying this chronological-based segmentation, we turn the problem of images retrieval into image segments retrieval, in which the boundary between activities such as having breakfast, working in front of a computer, and so on [43], are automatically decided based on the time and concepts. Starting from a topic query, it is transformed into small inquiries, where each of them is asking for a single piece of information of concepts, location, activity, and time. The moments that matched all of those requirements are returned as the retrieval results. In order to remove the non-relevant images, a filtering step is applied on the retrieved images, by removing blurred images that covered mainly by huge object or by the arms of the user. Finally, the images are diversified into clusters and the top images that close to center are selected for the summarization, which can be done automatically by following the methods in [143]. These steps are described as follows:

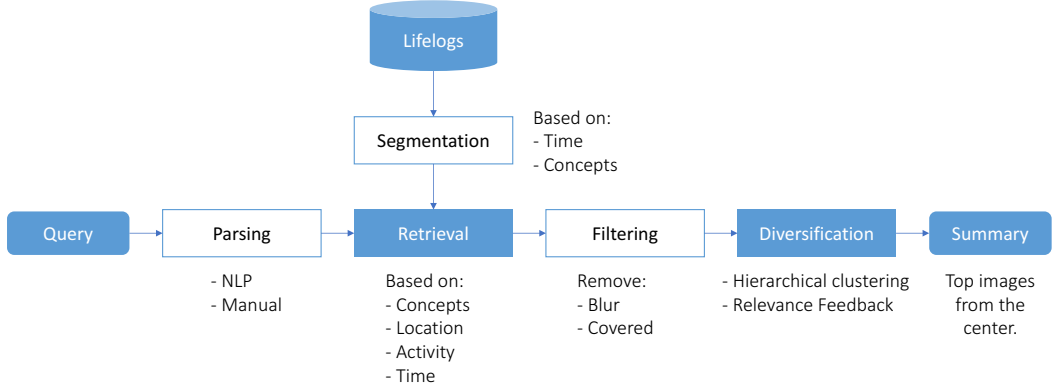


Figure 5.4: Schema of the proposed methods

## Segmentation

For the segmentation we applied a simple chronological-based segmentation as follow:  
For each pair of two continuous images  $I_t$  and  $I_{t+1}$  at the time  $t$ , the distance  $d(I_t, I_{t+1})$  between them is computed as:

$$d(I_t, I_{t+1}) = \|\vec{C}_t - \vec{C}_{t+1}\|$$

where  $\|\cdot\|$  is the normalized Euclidean distance, and  $\vec{C}$  is the concept vector of each image provided from the task. If  $d(I_t, I_{t+1}) < \tau$ , where  $\tau$  is a threshold, the two images are set belong to the same segment, otherwise they are set in different segments. If  $\tau$  is too small, an activity should be split into small activities, while larger value of  $\tau$  should be grouped different activities into the same one. Since  $\|\cdot\|$  is normalized, when  $\tau = 0$ , the images are grouped into different segments, and when  $\tau = 1$ , all images are belongs to a single segment.

Segmenting the activities is not simply an incident of identifying the exact event boundaries; it also concerns with keeping track of the fine-grained group of events together into extended meaningful units, and thus deciding the right value of  $\tau$  is not trivial. In the proposed approaches, we try different values of  $\tau$  for different runs.

After this step, each segment is represented by the first image (of that segment)

with these basic information: location, activity, time segment, number of people and the list of the concepts. If any of these information is missing from the first image, we take it from the second image and so on.

## Filtering

An image can be considered as blurred based on its focus level. In the proposed approaches, we estimate the focus by computing the absolute sum of the wavelet coefficients and comparing it to a threshold, by exploiting the method in [144]. The return of this method is a scalar number in  $[0, 99]$  which the bigger value the sharper image. From our observation, for values below 30, most of the images are blurred, and thus we set this threshold to 30.

In order to remove images that covered by large objects, we applied an heuristic method as follows:

Step 1 Convert the image to binary images by applying thresholding with several thresholds.

Step 2 Extract connected components and calculate their centers.

Step 3 Group centers based on their coordinates, and then close them to form the corresponding blob.

Step 4 Take the biggest blob and its size (in pixels).

If the size is over 50% of the whole area, the image is considered as covered. This whole method is implemented by calling the function `SimpleBlobDetector` from OpenCV<sup>1</sup>.

After this step, all remain images are considered as relevant to the topic. Please notice that the images are still kept inside the segment.

---

<sup>1</sup><http://opencv.org>

## Diversification

In this step, for automatic approach, we use a hierarchical agglomerative clustering algorithm (see in [145]) to group similar segments into the same cluster based on the concepts. The clusters are then sorted based on the number of segments, decreasingly. Finally, we produce the summary for the queries by selecting representative images from the clusters which are the images closest to the center of each cluster.

We also propose a human-in-the-loop approach in this step, that asks the user to assign the labels *Relevant* \ *Non-relevant* to the retrieved images. The system asks the user to label the representative images of the top  $N$  results returned by the automatic diversification procedure (as mentioned above), and the number of images that have been labeled as being *Relevant* \ *Non-relevant* for each cluster is computed. Then, the clusters are sorted as follows:

- Clusters that have a large number of relevant counts are sorted higher.
- Clusters that have the same number of relevant counts are sorted based on the number of non-relevant counts (i.e., a cluster that contains a larger number of ‘non-relevant’ images should be selected later).
- Clusters that have the same number of *Relevant* \ *Non-relevant* counts are sorted on the basis of the number of segments.

For each cluster, the images that are selected to represent the topic are chosen in the same way as in the automatic diversification.

## Submitted Runs and Results

We submitted 3 runs on the retrieval sub-task, summarized in Table 5.1. For the retrieval task, the first run is exploiting only time and the concepts information. We consider every single image as the basic unit and the retrieval just returns all images that contains the concepts extracted from the topics. We named this run



is the ‘baseline’ with the purpose that any other approaches should obtain better performance than this. With the second run, we applied the optimized value for  $\tau$  (optimized from the development set) to do the segmentation. So in this run, the only difference is the basic unit of retrieval, now is the segment, not image. For the fine-tuning runs, the “translation” which described in 5.2.2 is applied in retrieval steps.

Table 5.1: Submitted Runs.

RunID	Name	$\tau$	Parsing
LRT Run 1	Baseline	0	Automatic
LRT Run 2	Segmentation	0.05	Automatic
LRT Run 3	Fine-tuning	0.05	Fine-tuning

The Table 5.2 shows the result of the runs on the retrieval. The results confirm that applying segmentation improved retrieval. It is quite clear that applying fine-tuning significantly improved the performance. The big gaps in results between the automatic approach with the fine-tuning and between the fine-tuning with the human-in-the-loop approaches, has shown that we need better natural language processing as well as machine learning studies for these problems.

Table 5.2: Lifelog Retrieval Results.

Run	Name	Average NDCG
LRT Run 1	Baseline	0.09
LRT Run 2	Segmentation	0.14
LRT Run 3	Fine Tuning	0.39

### 5.3.2 Learning Improvement

In this section we introduced different baseline approaches, that came from fully automatic to fully manual paradigm. These approaches, that require different level of involvement of the users, exploit only the information provided by the organiser along with the collection of images, i.e., the description of the semantic locations and the physical activities. From the obtained results it appears clear that deeper analysis of the methods should be considered as well as the use of extra information

(e.g. relevance feedback). The relevance feedback about the temporally retrieved image, which leads to the development of system 3.

## 5.4 System3: LIFER

This section describes the participation of the Organiser Team in the ImageCLEFlifelog 2018 Lifelog Moment Retrieval. In this section, we propose how to exploit LIFER, an interactive lifelog search engine aiming to solve the Lifelog Moment Retrieval task. We propose approaches for baseline, which aim to provide a reference system for other approaches, and human-in-the-loop, which advance the baseline results.

### 5.4.1 Interactive Lifelog Search Engine

Our proposed solutions for the Lifelog Moment Retrieval task is to develop a new the baseline system, LIFER [146], which is an interactive engine for lifelog retrieval. LIFER is improved upon an existing baseline search engine described in section 5.3, which was developed to provide a starting point for researchers engaged in collaborative benchmarking exercises, such as NTCIR [33] and this ImageCLEFlifelog2018 [59] tasks. It was also used for the LSC@ICMR 2018 [57] competition and got a reasonable result. In this section, we will introduce how we used the LIFER system to address the LMRT task and list the approaches we used for retrieval images information.

LIFER uses the core search engine I developed in [142] which offers a platform for searching images that match with some criteria. This interactive system helps user to retrieve results in a faster and reliable way, which helps to solve LMRT sub-task. The detailed operation will be described in Section 5.4.2. LIFER is built based on the six sources of information which was extracted from ImageCLEF dataset.

- Time. The most basic unit of data in the dataset, time gave us the possibility of including more semantic concepts, such as days of the week, weekday/weekend,

times of the day, etc. In the LIFER system, we consider the unit of time as minute, i.e., each image is attached to a minute. These time is extracted (and linked to the image) directly from the provided data.

- **Locations.** Semantic location were provided in the dataset which provided localised names for all visited location. For example ‘The Helix’, ‘Dunnes stores’, ‘Dublin City University’ and so on.
- **Visual Concepts.** Enhanced visual concepts extracted by Microsoft API [147] was provided, which accompany with each image. These visual concepts were indexed in our lifelog retrieval system. Visual concepts describe the content of the lifelog images included in the dataset. Each image has one or more concepts identified and tagged. The concepts (in text form) were indexed.
- **User Activities.** The physical activities of the user (e.g. walking, sitting, running, etc.) were indexed as additional search terms.
- **Biometrics.** The biometrics of the user were also indexed as semantic labels. These included the Galvanic Skin Response (stressed/excited, relaxed) which can be considered to be a correlate of stress or excitement levels, and the level of physical activity (exercising / resting) as identified from the heart rate.
- **Music.** A log of the music listing history of the lifelogger was included in the collection and we considered that it could be an important aspect of some topics. The song\_name and song\_artist are two options which are used in search results.

These six sources of information are instantiated in the user interface as facets of a user query, as shown in Figure 5.5.

The Interface of LIFER is shown in Figure 5.6. The upper section of the interface is the query-panel in which the faceted queries are created. Below that is the main part of the interface for displaying the selected lifelog images in temporal sequence.

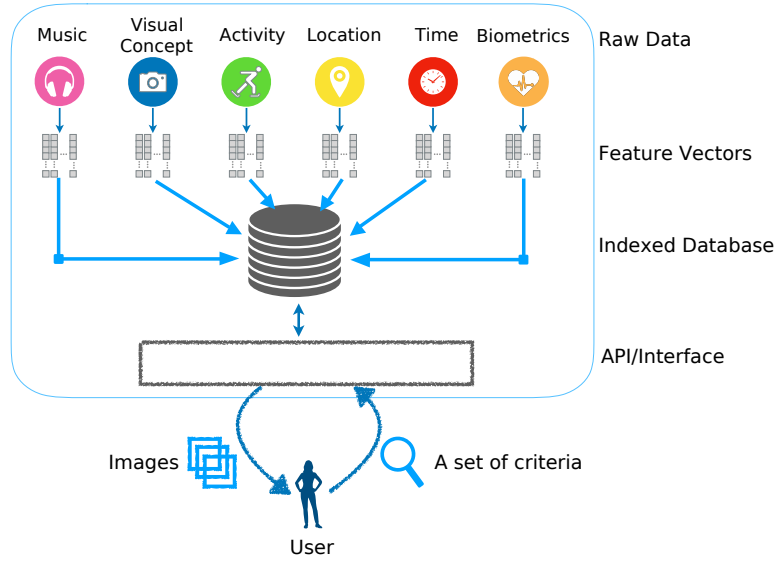


Figure 5.5: Schema of LIFER, the proposed interactive lifelog search engine.

LIFER: Interactive Lifelog Retrieval System

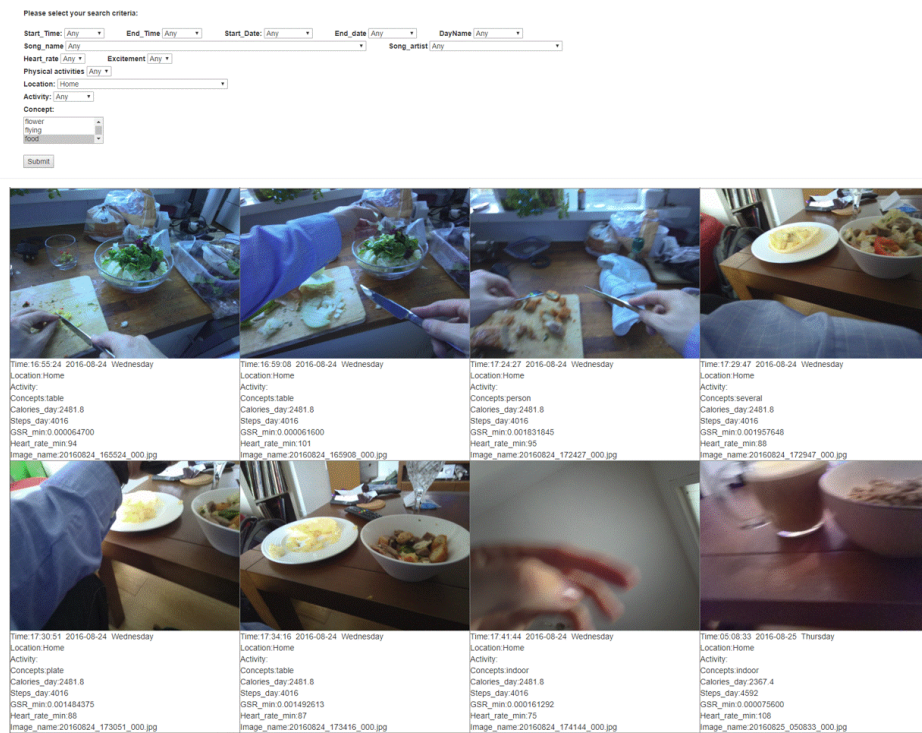


Figure 5.6: The Interface of LIFER with an example on the results for solving a topic of "Find the moment the use is having vegetables".

The URL for LIFER is <http://searchlifelog.computing.dcu.ie>

In the query-panel, the search facets are shown. The facets are directly related to the indexed data (see the six sources of information above). Upon submission of a faceted query, the system returns a temporally organised listing of potentially relevant images. In the first version of LIFER, the query facets are combined in an AND boolean manner and it can be changed on a per-topic basis.

The temporally organised listing of relevant images are displayed in the lower part of the screen (the result-display panel). Each relevant image is listed with an overview metadata as a form of context. This metadata is configurable to display various sources of information, as required. Figure 5.6 shows a basic form of such metadata.

#### **5.4.2 Using LIFER for ImageCLEFlifelogs2018 Retrieval Task**

As mentioned, we used LIFER for ImageCLEFlifelogs2018 LMRT task. Firstly, based on the topic description, the search criteria are determined (by both automatically considering all the words in the queried topic as concepts, or alternatively allowing concepts to be determined by the user). Secondly, we applied simple relevance feedback mechanisms in LIFER by using the usual standard Relevance Feedback paradigm (more details can be seen in [143]) to allow user manually select multiple relevant images for the submission or taking all of them as relevant. In term of ranking, we determine to use the default option of LIFER by chronological order.

We submitted 3 runs for LMRT subtask, which are summarized in Table 5.3. For LMRT, the first two runs are automatic while the other three adopted the human-in-loop. Tables 5.4 show the search criteria for the best run of each task. Since the existing gap between the abstract understanding of human needs and the multi-modal data, the best results are still coming from the fine-tuned queries.

Table 5.3: Submitted Runs for LMRT.

RunID	Name	F1@10	Notes
LMRT Run 1	Baseline	0.077	Fully automatic without ranking
LMRT Run 2	Baseline	0.131	Fully automatic with ranking
LMRT Run 3*	Baseline	<b>0.407</b>	Search by main terms and Human filtering

\* These runs were submitted after the competition.

Table 5.4: Selected Criteria for LMRT Task.

Topic	User	Activities	Times	Locations	Concepts
T001	u1	-Running, -Transport	+MinuteID: 540-1140(workday)	+Work, -Home	+Coffee, +Cup
T002	u1	-Running, -Transport	+MinuteID: 400-540(15/08/2016 to 15/09/2016)	-Work, -Home	+Shop, +Store
T003	u1	-Transport, -Airplane	+MinuteID: 400-1400	-Work, +Home	+food, +vegetables, + Kitchen
T004	u1	-Transport, -Airplane	+MinuteID: 400-540(01/09/2016 to 30/09/2016)	+Home	+TV
T005	u2	-Waking, -Running, -Transport	+MinuteID: 400-1400(01/09/2016 to 30/09/2016)	+Work, -Home	+People, +Indoor

+ means selection and – means exception.

### 5.4.3 Learning Improvement

In this section we introduced an interactive lifelog retrieval system LIFER to tackle the ImageCLEFlifelog 2018 task, as a participant of the LMRT subtask. These approaches, that require different levels of involvement from the users, exploit only the information provided by the organisers along with the collection of images, e.g., the description of the semantic locations and the physical activities. As we can see in Table 5.3, without the manually input, the results are significantly worse (as in the result of LMRT Run 1). This shows that the key challenge is how to translate the query to the search criteria, with requires further study.

There were a number of limitations to the LIFER search engine. The query

panel was not optimised in terms of layout, the result list of images is currently only temporally sorted. The metadata highlighted could be query-specific and the metadata itself should be capable of generating a query to the system, so as to facilitate a richer browsing methodology. In order to optimize LIFER, we plan to upgrade the interface using the facets to query which support this temporal sort function. Additionally, for the purpose of bridging the gap between the query understanding of human needs and the multi-modal data, we choose to utilize the state-of-art Faster R-CNN [148] technique to detect objects. These are the learning improvements for the system<sup>3</sup>.

## **5.5 System4: LIFER2.0**

I worked with team partners to propose and develop the LIFER2.0 lifelog retrieval system based on learning from previous three system. This section describes our participation of the Organiser Team in the ImageCLEFlifelog 2019 Lifelog Moment Retrieval (LMRT) tasks. We used LIFER 2.0, an enhanced version of LIFER, which was an interactive retrieval system for personal lifelog data designed for novice user application. Considering the multi-modality of lifelog data, we pre-processed the images to remove noisy data as a first step and then focused on the exploitation of associated metadata (time, activities, location, etc.) from moments of daily life. Inheriting the structure of the interactive search engine from [146], we developed a new facet filter and context browsing interface, with additional visual concepts and criteria expansion for ImageCLEF2019 LMRT. The results on LMRT task confirmed that by using faceted filter and context browsing, a user can gain insights from their personal lifelog by employing very simple interactions.

### **5.5.1 Retrieval**

For ImageCLEFlifelog2019 LMRT task, we used LIFER 2.0 - baseline interactive search engine which was initiated in [146], and gained the improvement on object

detection and flexibility of interface operation. In this section, we provide a description of the interactive retrieval system and how it can be used to solve information needs. LIFER 2.0 system, as described in [78], is a criteria matching engine which is built mainly from five categories: date/time, location, activity, biometrics, and visual concepts.

1. **Date/Time:** Date/time is an important feature in LIFER 2.0 search engine system because it can narrow down the scope of moment searching. For instance, time is specifically useful in query 6: "Having breakfast at home" (must have breakfast at home from 5:00 am to 9:00 am). It could also be useful for result filtering and lifelogger's behaviour guessing. In this system, date/time criteria include week days, date, and time.
2. **Location:** Location criteria contain location categories and location names, which are also advantageous for users to retrieve the relevant images in topic 1, 5, and 6. These topics depend mostly on location filtering to find the proper moments and increase the variety of the chosen images.
3. **Activity:** Although activity metadata in ImageCLEFlifelog 2019 dataset is not diverse, it is a potential criterion to be integrated into LIFER 2.0 system to improve the search engine with user actions/behaviours when it is ready.
4. **Biometrics:** Due to the lack of activity information, biometric data provide us the means to guess the moments when lifelogger is eating, walking, moving by heart rate and calories changes.
5. **Visual Concepts:** These concepts play the key roles in finding the proper images for topics owing to the diversity of concepts, annotations, and keywords. They include place attributes, place categories, and objects' name. Place attributes and categories are extracted from places365-CNN [149] with top 10 extracted attributes and top 5 place category predictions. Objects in image are detected using Faster R-CNN [148] trained on MSCOCO



dataset [150].

The interface of LIFER 2.0 system was divided into two parts: facets filter and context browsing. For the facets filter, a user could adjust his/her choice of five aforementioned criteria to retrieve the desired moments. In each criterion, except for the location, the keywords and tags are combined into a query condition using the OR operator to expand the diversity of returned results. Finally, all the conditions from each criterion are merged into a final query by utilising the AND operator. For context browsing, the keywords and annotations from location, visual concepts, activity are added into an auto-complete search bar. The user then types and chooses tags which are suitable for current context of each topic. The query processing of this function is the same as the facets filter. The interface of LIFER 2.0 - baseline interactive search engine is demonstrated in Figure 5.7.

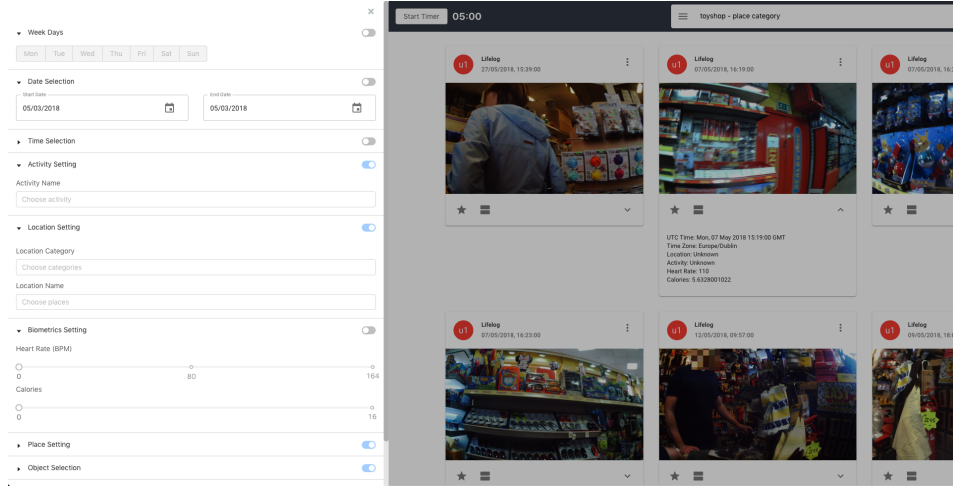


Figure 5.7: The facets filter (left) and context browsing interface (right) of LIFER-baseline interactive search engine with an example of shopping in toyshop.

### 5.5.2 Experimental Configuration: LIFER 2.0

For benchmarking systems, we provided an efficient retrieval system, based primarily on faceted querying using captured metadata, served as a baseline for other systems, and provided the basis for the LIFER 2.0 system.

## LMRT Task Results

LMRT was an ad-hoc retrieval task and Table 5.5 displays the result of our two runs as the participant. As can be seen from the table, we achieved 0.41 in terms of precision, with cluster recall of 0.31 and 0.29 in F1 score. Figure 5.8 and Figure 5.9 give us a precise look into multiple cut-off positions of the returned ranking for each query of both runs. We observe that the system has its stability across users as both graphs share the same pattern over three metrics. In LIFER2.0, the analysis demonstrates that the search engine increased the F1 score by increasing cluster recall through valid experiment criteria.

Table 5.5: Submitted Runs for LMRT task.

RunID	P@10	CR@10	F1@10
LMRT Run 1	<b>0.41</b>	<b>0.31</b>	<b>0.29</b>
LMRT Run 2	0.33	0.26	0.24

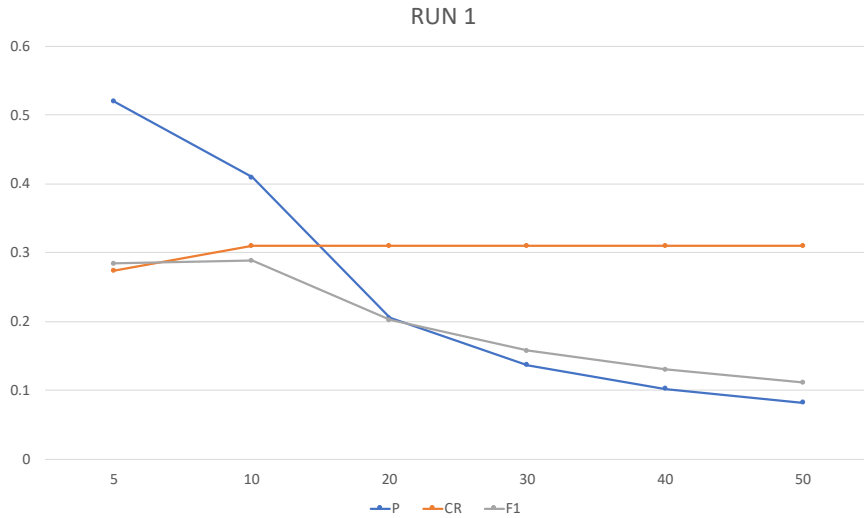


Figure 5.8: Result of Run 1 in various cut-off positions

## 5.6 Conclusion & Contribution

In this chapter, we address the research question of applying traditional information retrieval techniques on this multi-modal lifelog data to develop a high-quality baseline

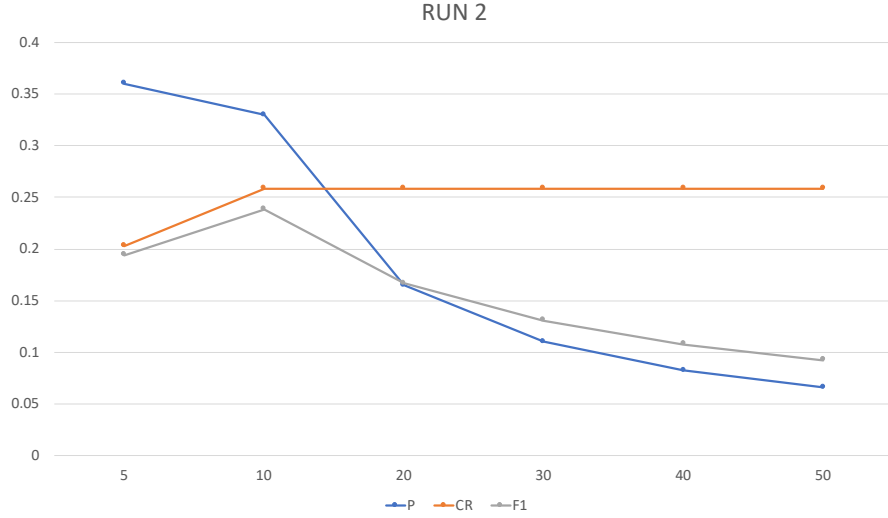


Figure 5.9: Result of Run 2 in various cut-off positions

search engine. We are motivated to carry out this research, since the apparent lack of baseline search engine to help non-expert users to access large lifelog images content and serve as a baseline for other lifelog retrieval systems. At the end of this research, we aim to present an answer to our second research question which was defined earlier.

For the RQ2, we developed the baseline search engine for personal life archives, which aims to make the personal life archive searchable, organisable and easy to be updated. We also compared it with other approaches and get some preliminary results, which illustrate the feasibility of the baseline search engine as a tool for getting insights from personal life archives. In order to follow the action research approach, we also constantly tested our baseline approaches, from fully interactive which need human involvement to LIFER and LIFER2 system which are upgraded based on previous systems. This lifelog retrieval system was improved through these 4 cycles over 3 years using the different test collections, a direct comparison is not provided, but the performance enhancements are clear. We used the these retrieval systems for comparison in different comparative benchmarking initiatives and it has been available to use via API for other participants to use. These interactive retrieval systems were developed to provide a starting point for

researchers in the area and as a baseline for our future work. Besides, through reviewing the work from other participants, we gradually learned that multi-modal data analysis should be explored and exploited, with the majority of the approaches combining visual, textual, location and other information to solve the task. Furthermore, we learned that many approaches are based on deep neural networks, from standard CNNs to specifically designed deep networks for lifelogging tasks. However, there are still rooms for improvement, since the best results are coming from fine-tuned queries, which means we need more advanced techniques to bridging the gap between the abstract understanding of human needs and the multi-modal data. Furthermore, automatically “translating” the query into the retrieval criteria is still a challenge which requires further study.

We developed and upgraded the baseline search engine for personal life archives in this chapter and contributions has been made in our work as following:

- We proposed and defined the first high-quality lifelog retrieval basedline search engine on standard IR approach.
- We participated different campaigns and upgraded the lifelog baseline system using the techniques including event segmentation, relevant feedback, object detection.

## 5.7 Chapter Summary

After examining these approaches and finding them to be useful, we investigate methods to answer our second research question. We demonstrate that we can address the research problem of applying traditional information retrieval techniques on this multi-modal lifelog data to develop a high-quality baseline search engine on a large lifelog collection. In order to build scalable lifelog search engines, our further work will focus on seeking the existing of semantic relations between different human actions existing in lifelog visual data and exploiting the importance of such semantic relations in lifelog moment retrieval task. The next chapter describes the

proposed methods for automatically lifelog image retrieval using enriched Human Object Interaction and causality features.

## Chapter 6

# Causality for Related Query Retrieval with Deep Learned Human-Object Interactions

### 6.1 Introduction

In this chapter we address the research question 3, which is **can we improve retrieval performance by enriching the standard visual lifelog features by integrating causality-inspired features to provide a state-of-the-art retrieval system?**

To answer this research question, we start our research by investigating and enriching the image concepts features using human object interaction(HOI) and the causality score between different HOIs, and we then compare the automatic and manual HOI features to validate which feature is vital in improving the accuracy of lifelog moment retrieval.

With the ever-increasing number of wearable cameras around us, there is a growing need for searching and inferring the context of captured events and the potential relationship that may exist among humans, objects, and their interactions. Despite of many advancements in computer vision such as improved instance and human action recognition, lifelog multimedia data understanding and prediction still remains a challenging task. In order to bridge the gap between the semantic understanding and vision representation, it is necessary to understand the

causality [151] between different human and object interactions. So we pioneer an approach to retrieval that integration causality.

Causal reasoning is the process of identifying causality: the relationship between a cause and its effect. Humans can reason about many topics with the aid of causal understanding and this naturally translates into semantic activity-based queries in retrieval systems. For instance, consider a security professional who is interested in finding surveillance video that shows a stolen item (e.g. handbag) being passed from one individual to another, or in other video retrieval applications (e.g. sports) where a human activity is a topic of interest, or even in personal data analytic (e.g. lifelogging) when a person wishes to query past activities captured using wearable cameras.

We aim to improve the performance by applying machine learning to enforce the attributes of Human Object Interaction, in particular with causality between different HOIs, carried out at two mutually correlated levels: feature level and ranking level. The former includes capturing the intrinsic interaction information on features by effective features transformation (feature selection, feature weighting, and feature mapping) and the latter aims to build effective ranking models for encoding the underlying structural ranking relationships among data samples. We will focus on the features weighting and mapping based on queries, and design a effective rank scheme to exploit needed variables.

Chapter 4 shows the basic benchmark retrieval system for lifelog information retrieval. The main research is built on concepts extracted and the metadata recorded. We believe that for some queries the semantic relationship inside the query can only be expressed by some interaction keywords. For instance, the query is "please find the moment the user is holding a cup to drink the coffee", it is obvious there is a human action "hold" or "take" and it is more complex than basic concept extracted and automatically annotated using a computer vision annotation. Even though a query "Find the moment a man or a girl give a book to another man ", we know that the moment should include the entire procedure of

“giving”, not only the moment of “giving”. It should also contain the moment of both man “holding and giving” in that period of time. But often the “giving” behavior is too fast to record using wearable camera. How to deal with this kind of potential interactive behavior between lifelogger and humans is a big challenge in our lifelog retrieval research. In order to motivate the need for this, there are 17/23 retrieval topics from LSAT task are analyzed in Table 6.1, which show whether some possible relationship exists between topic query with other HOIs.

Table 6.1: LSAT topics in NTCIR-13 Lifelog subtask. The causal means whether the relationship is existing with other HOIs

Topic Title	Casual
Eating Lunch	Yes
Fruit or Vegetable Juice	Yes
Working Late	Yes
Gardening	Yes
Photo of the Sea	Yes
Exercises	Yes
Castle at Night	No
Having Beers in a Bar	Yes
Use the Computer	Yes
Have Coffee	Yes
Greek Amphitheatre	No
Benbulbin Mountain	No
Sunset	No
Television Recording	Yes
Cooking	Yes
Graveyard	No
Working in a Coffee Shop	Yes
Hiking	Yes
Presenting / Lecturing	No
Painting Walls	Yes
Flying	Yes
Grocery Shopping	Yes
Eating Pasta	Yes

There is a semantic relation that possibly exists between different human behaviors, which can be viewed as *Causality*. In order to address the research question 3, we need to know whether we can extract human object interaction and infer the quantified relationship to enhance retrieval. In this chapter we explain our



methods to enhance the performance of lifelog retrieval by taking causality into account.

As we know, notwithstanding recent advances in machine vision, visual activity recognition still remains a challenging task as many real-world interactions cannot be automatically recognised for many reasons, such as partial occlusion or coverage black-spots. In this chapter, we propose a new technique that infers the unseen relationship between individuals' behavior captured visually and use it to retrieve relevant lifelog moments if there is a likely relationship between individuals' different behavior. We introduce a Human-Object Interactions (HOI) model integrating the causal relationship between the humans and the objects. We will provide detailed information about all the steps in the next section.

A person has multiple interactive actions and there are some inherent relationships existing between people and object. So establishing causal relations between variables from observation of their behaviour in time is central to scientific investigation and it is at the core of data-science where these causal relations are the basis for the construction of useful models and tools capable of prediction which can be used for related human behavior retrieval.

Learning causality has been applied in myriad of high-impact applications that include computer vision [152, 153], financial data analysis [154, 155], environmental health [156, 157] and education [158, 159]. Also, causality has been viewed as a significant component to improve the accuracy of information prediction and retrieval [61, 62]. Many approaches in recent times combine machine learning with causal inference [160] to answer causal questions [152] and to build a causal inference framework to answer questions regarding the causal relationships between object statuses and human actions, which in many cases may be hidden or not visible.

In this chapter, we report the performance of a first-generation approach to solve this challenge, demonstrating that the proposed model can successfully infer causality between the different human-object interactions in video and lifelog data.

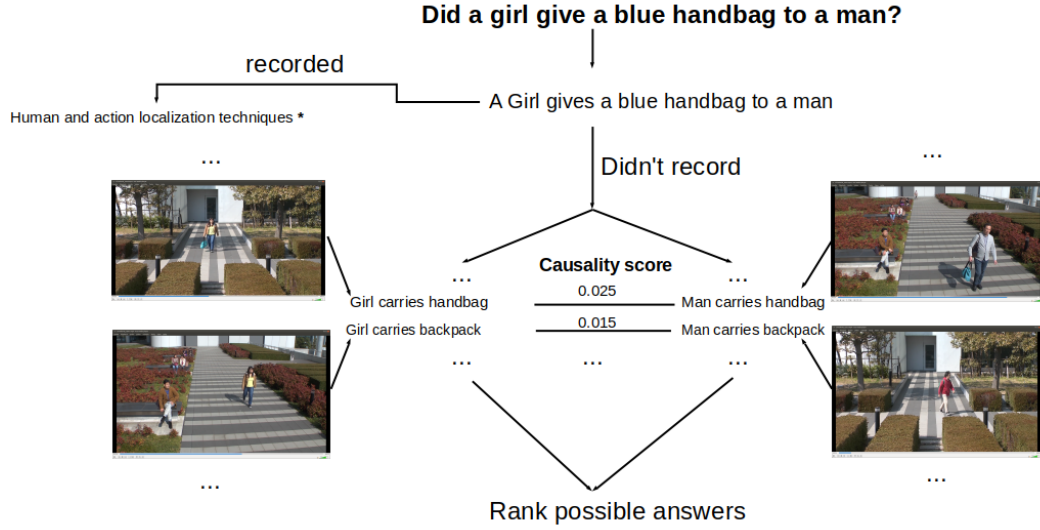


Figure 6.1: An example of the challenge that shows a stolen item (e.g. handbag) being passed from one individual to another: two individuals carrying the same blue backpack, with a probable handover of the backpack not captured by the surveillance cameras.

Figure 6.1 summarises the research challenge visually, which shows two individuals carrying the same book or bag, with a probable handover of the backpack not captured by the wearable or surveillance cameras.

## 6.2 Human Object Interaction Extraction

There is some research on HOI detection and recognition in image data. Human object interactions [161] are related to visual relationships, but present different challenges. HOI [161] is a necessary attributes for lifelog information retrieval. For most queries, it is not only includes the concepts and attributes, but also the human action or human object interaction. It is more challenging to analyse and recognise the human object interaction. Because the contextual and semantic understanding of lifelog visual dataset is a big challenge in computer vision. Detecting a Human-Object Interactions (HOI) can provide deeper insights to understand the semantic meaning of the lifelog dataset. Some research work has been applied on Human-Object Interactions (HOI) detection and recognition.

Chao [128] used the CNN to encode object instance appearances and the relative spatial relationships between a person and the object. Zhou et al. [162] proposed a cascade network architecture which comprises instance localization network and an interaction recognition network for HOI relation ranking and classification. Xu et al. [129] improved the HOI detection model in social scene images to combine human pose and gaze estimation together.

After reviewing the literature and the related research, we found the existing work recognizes interactions based on human or objects cues (either human appearance, object appearance, or spatial relationship between human-object pairs). It seems such predictions inevitably lack contextual information. Meanwhile, Gao et al. [130] proposed instance-centric attention module which uses attention mechanism to extract contextual features to facilitate HOI detection. The ICAN model can learn to dynamically highlight regions in an image conditioned on the appearance of each instance. Besides, the human actions are fine-grained (e.g., walking, running, surfing, snowboarding) than the normal actions of general subjects, and an individual person can simultaneously take multiple actions (e.g., drinking tea and reading a newspaper while sitting in a chair). This requires a deeper understanding of human actions which can affect the HOI retrieval results. After comparing with other related work, we found the iCAN [130] model supports and boosts the HOI recognition performance via integrating the contextual information. So we choose to build upon the Gao et al. work to detect the human object interaction to extract human object interaction triplets of video and lifelog images data.

### 6.3 Causality Scenario in Lifelog and Security Data

We conducted an initial investigation in a well understood domain (surveillance video) which simulates the lifelog issue of physical missing data. Surveillance video data also exhibits the “missing data” problem. The detailed scenario about this

problem in security and lifelog data are described below.

### Scenario in Security Data

In order to find an initial solution to the problem of causality detection, we present a short scenario in the domain of lifelog images and captured security videos, with a typical layout of security cameras. In an outdoor environment, there is a U-shaped pathway which is configured with 3 cameras: 1, 2 and 3, as shown in Figure 6.2. As can be seen, some corners (bottom center) are not covered by the cameras, and are in effect hidden regions, which we refer to as 'dead-zones'. While cameras 1, 2 and 3 are separately recording their own field-of-view, they will not capture activities in the dead-zones. Consider a scenario in which there is a person (A) carrying a blue handbag, which is given to a person (B) in a dead-zone (i.e. not captured by any surveillance camera). Camera 3 captures footage of a person carrying a similar blue handbag (the same bag) that was carried by person A, a few moments previously. Without causality detection, this handover is not identified by the security system, which could potentially pose a major security risk. An information need such as "*did a person give a blue handbag to another?*" can not be answered without a human-object interaction causality model. Our challenge, in this work, is to find and infer causality between unseen actions, based on the video captured from two or more different cameras.

### Scenario in Lifelog Data

In lifelog moment retrieval domain, there are some similar scenarios like "Lifelogger give a book to a man", there is a potential relationship between human behavior. Lifelog images dataset contains temporal images and the approaches proposed for video also can be applied in lifelog research.

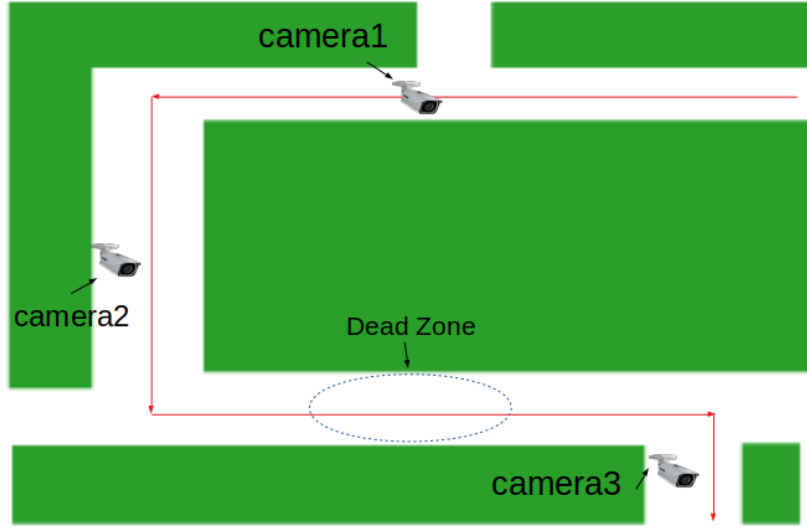


Figure 6.2: Scenario Map showing “dead zones”.

### Analysis on Scenario

In order to tackle the problem highlighted by the above scenario, we propose a HOI (Human-Object Interaction) causality model to infer unseen interactions. The overview of the HOI causality module is shown in the Figure 6.3, which is fully designed by the dissertation author and we use as the basis for the following explanation of the model. The framework is a pipeline consisting of four components: keyframe extraction, SVO extraction, encoder & HOI model, HOI inference. Note that in this proposed approach, we select state-of-the-art components, but further research could be focused on enhancing the performance of each component, thereby enhancing the performance of the overall technique.

## 6.4 Human-object Interaction Causality Model for Video and Lifelog Retrieval

The proposed HOI causality model operates by identifying human actions and contextual features from video and lifelog images. The process can be described as follows: firstly extract the instance-level appearance features such as object colour,

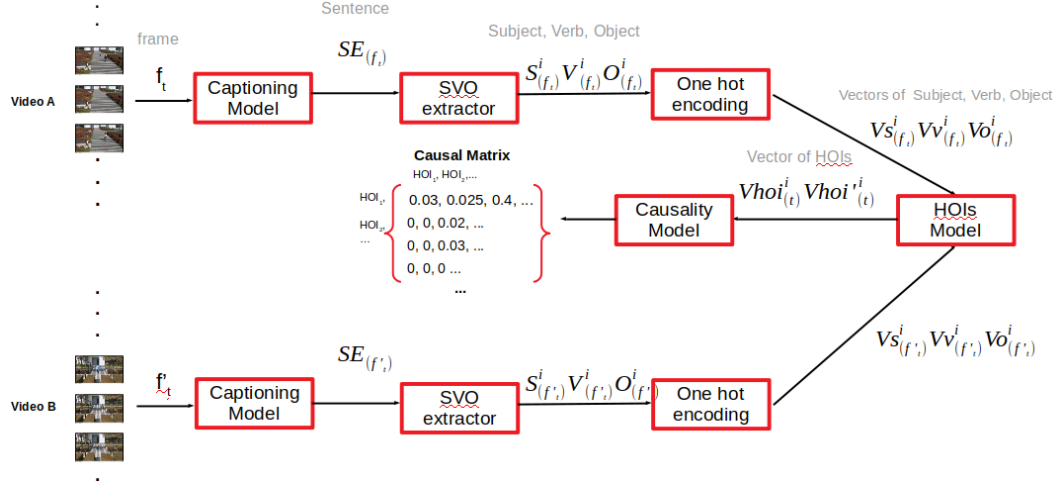


Figure 6.3: A brief illustration of our proposed model for HOI causality

shape, and gender using a standard process to distinguish humans and objects. Secondly, with a goal to dynamically generate a causality map conditioned on the different HOIs corresponding video frames and lifeloga images are retrieved from the dataset based on a generated causality matrix, which is a graph encoding interactions between known objects in the images. The details of developed techniques are explained in the following sections. We provide an overview of the entire approach in Figure 6.3 and outline the method for the inference in section 6.4.4.

### 6.4.1 HOI Recognition

In order to better understand and extract HOI, we tried two methods to extract HOI triplet which will be used to calculate the causality score. One is built based on LSTM-Image-Captioning model, the another is using ICAN Model.

#### Image Captioning Model

Image captioning is the process to generate textual description of the input images. Currently, automatic image and video captioning is extremely important, as well as challenging, because it connects two major artificial intelligence fields: computer vision and natural language processing. Previous work [163] use recurrent neural

networks which is based on long short-term memory units as decoder. In our work we choose this model to extract sentence which is used to extract the HOI triplets.

For this, we first annotate images automatically using the state-of-the-art image captioning models. Next, we extract SVO (subject, verb, object) triples and encode the descriptions into a vector form for HOI inference using the Stanford CoreNLP parser (<https://stanfordnlp.github.io/CoreNLP/>). In order to calculate the HOI co-existence and the possible causality score we use transfer entropy. From our experimentation, we found that integrating casual relations into the content indexing process and using transfer entropy to calculate the causality score leads to improvement in retrieval performance.

The second phase of the process (see Figure 6.4) is image captioning, which generates labels or annotations for the previously extracted key frames  $SE_f$ . We consider the problem of HOI generation as being similar to the problem of video captioning. Currently, the state-of-the-art video captioning results are based on using LSTM and CNNs. In our implementation, we use CNNs to extract image features and a multiple LSTM structure to extract the main image caption features and employ multi-layer LSTMs with 512 dimensional hidden states to model the probability of different words in the captions. The reason to choose this captioning model is that we want to use a multi-point process to qualify the existence of HOIs and apply this multi-point process in causality transfer theory. Hence, using the image captioning models is a pre-processing step in our method.

The inputs to the HOI generation process are two surveillance videos or a period of lifelog moments captured using different cameras that are located in different locations, but within close proximity to one another, and with a dead-zone between their fields-of-view. The output of the captioning model is a vector of sentences describing the key-frames in terms of objects and activities (along with their scores) from each video sequence. Like the scenario described earlier in which two people are seen to be holding the same blue bag in different camera sequences. The challenge to be addressed now is how to infer the reasons why the bag has been transferred from

one person to another; was the bag handed over, or was it potentially stolen? This corresponds to the SVO extraction process that is explained in the next section. An example of the image caption for a key-frame based on image captioning model is also given in next section.

$$SE_{f_t} = \text{argmax}(CPM(f_t)) \quad (6.1)$$

The input of the image caption module (labeled CPM in formula equ7) is two images from different clusters. The key images from different clusters: Input1:  $f_t = \{v_1, v_2, v_3, \dots, v_n\}$ , Input2:  $f_t = \{v_1, v_2, v_3, \dots, v_n\}$ . The output SE (SE stands for the image caption sentence) of image captioning is a sentence description of the input frames. The example of the sentence is shown in the Figure6.4.

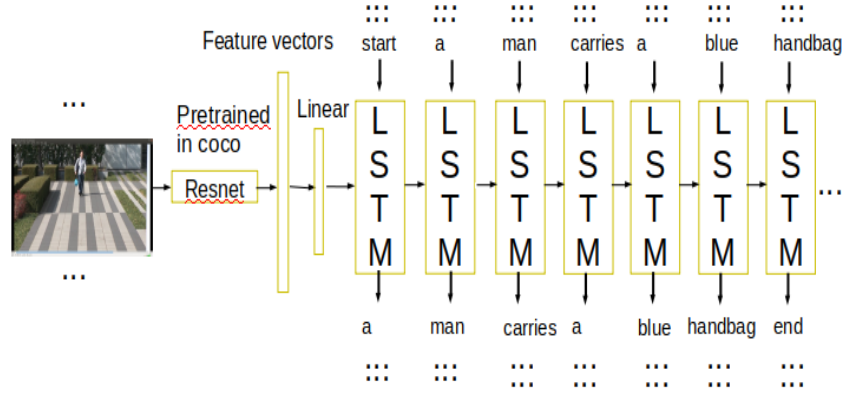


Figure 6.4: Image caption model

## ICAN Model

To investigate if there is improvement in the performance, another tool is also implemented to extract Human-object interaction automatically using iCAN model which is designed based on iCAN model [130] and an instance-centric attention network for detecting and recognising the interaction between each person and the object they are interacting with. An example of the HOI generated for a keyframe (using the iCAN model descried later in the section) is shown in Figure 6.5.





Figure 6.5: HOI output example for one keyframe from video showing detected object instances and recognized the interactions between person and object they are interacting with (carry handbag).

#### 6.4.2 SVO Extraction

SVO stands for subject, verb, object triplets ( $\langle S, V, O \rangle$ ) that are extracted from the queries and image captions. After generating the descriptions for videos in the previous HOI generation process, we analyze these descriptions via HOI reasoning. The first step is to extract the triplets that will be used to calculate the causality from these descriptions. We choose to use the state-of-the-art Stanford natural language parser [164] created by Stanford CoreNLP for this task, which generates an appropriate grammatical structure of sentences, for instance, which groups of words go together (as “phrases”) and which words are the subject or object of a verb. Given an input image as shown in Figure 6.3, the output of this SVO Extraction process would be a sentence in the form “a man is walking on the sidewalk with a blue bag in hand”. The next phase (shown on the right side of Figure 6.3) is to encode the textual descriptions from SVO extraction into a vector form for HOI inference).

### 6.4.3 HOI Model

The input for the causality model must be a sequence of time-series data, hence we have to encode the  $\langle S, V, O \rangle$  triplets into vectors which can be used to represent the presence of a Subject, Action, and Object triplet. The chosen encoder method is one-hot encoding, which is one type of vector representation of text for machine learning. The one-hot encoding is a representation of categorical variables as binary vectors and each word value is represented as binary vector that encodes zero values for all items except the index of the chosen word.

After extracting the binary vectors from the SVO, we employ the HOI model to combine the presence of Subject, Action, and Object triplets. The combination of the same HOI at nearby times extracted from different videos, is the main novelty of our HOI model. Given video A captured by camera 1 from time  $t_s$  to  $t_e$ , and video B captured by camera 2  $t_s$  to  $t_e$ , the challenge is how to combine these into HOI models. The detail of how to process the combination is concatenating the vectors from the two videos sequences during the same time period and adding them together when the subjects, actions, or objects are same.

**Problem:** Given the video A captured by camera 1 from time  $t_s$  to  $t_e$ , and video B captured by camera 2  $t_s$  to  $t_e$ .

In order to produce the temporal series data about HOI, we label people using different number to separate the HOI caused by different people. Subject:  $S = \{s_1, s_2, s_3, \dots, s_n\}$ , Verb:  $V = \{v_1, v_2, v_3, \dots, v_n\}$ , Object:  $O = \{o_1, o_2, o_3, \dots, o_n\}$ , Time:  $T = \{t_0, t_1, t_2, t_3, \dots, t_n\}$ . In video A, individual A appeared at time  $t_1$  holding a blue handbag and Individual B appeared at time  $t_2$  without any handbag. In video B, Individual A appeared at time  $t_3$  without a blue handbag and Individual B appeared at time  $t_4$  with a blue handbag. The comparison of time is  $t_1 < t_2 < t_3 < t_4 \subset \mathbf{T}$ . We argue that the combination can be formulated as follows:

$$A : Vs_{ft1}^i + Vv_{ft1}^i + Vo_{ft1}^i \quad (6.2)$$

$$Vs_{ft2}^j + Vv_{ft2}^j + Vo_{ft2}^j \quad (6.3)$$

$$B : Vs_{ft3}^i + Vv_{ft3}^i + Vo_{ft3}^i \quad (6.4)$$

$$Vs_{ft4}^j + Vv_{ft4}^j + Vo_{ft4}^j \quad (6.5)$$

The formula above represents the operation to add the SVO. The formula below combines the SVO from video(Moment) A and video/Moment B.  $K=\{s,v,o\}$ ,  $t$  and  $t'$  means the frames which contain  $S^i$  and  $S^j$ .  $S$  means the sum of different vector.

$$S(A, B)^i = \sum_{k,t} (Vk_t^i + Vk_{t'}^j) \quad (6.6)$$

#### 6.4.4 HOI Inference

HOI inference part is concerned with how to quantify relationship between different HOI triplets which are in the form:  $\langle Subject, Interaction, Object \rangle$ , extracted from the image captions. In order to calculate the level of relationship between different HOI values, we utilise a method named transfer entropy [165] to count the co-existence of Subject, Object, Actions and then compute the possible causality score between them. The causality of the time dimension imposes that past actions have consequences for subsequent actions or events. While the human and object appearance features contain strong cues for recognising the interaction, using appearance features alone often leads to possible inference results being missed.

#### Non-parametric Transfer Entropy

Wiener's causality theorem [166] states that one variable could be regarded as the cause of another if the ability to predict the future of the second variable is enhanced by utilising information concerning the first variable. Hence, we consider

to apply transfer entropy to infer the relationship between different HOIs and we create the causality score matrix (labeled CSM in formula 6.7) to show the causality between them. We use transfer entropy and obtain a scalar causal score from each causal measure by integrating over the frequency of occurrence. As we described in section 6.4.3, the decoder we choose is one-hot encoding, which uses binary labels to denote the existence of the human Subject, Actions and Objects. After processed by the decoder, the vector will be transferred into the HOI model, which adds the vectors to the other HOI vectors extracted from different video to produce the relatively complete HOI vector. For example:  $I = \langle x_1, x_2, \dots, x_n, \dots \rangle$ ,  $J = \langle y_1, y_2, \dots, y_n, \dots \rangle$ . The causality model will calculate the transfer entropy score using formula 6.8 and create the causality matrix.

$$S_{TE} = \operatorname{argmax}(CSM(HOI_i, HOI_j)) \quad (6.7)$$

The detailed *transfer entropy* is defined as:

$$TE_{HOI_j \rightarrow HOI_i} = \sum_{x_{n+1}, x_n, y_n} p(x_{n+1}, x_n, y_n) \log \left( \frac{p(x_{n+1} | x_n, y_n)}{p(x_{n+1} | x_n)} \right) \quad (6.8)$$

## 6.5 HOI Annotation

In order to compare whether the HOI and the concept of causality is useful in lifelog information retrieval, we compare our approaches with the automatic and manual annotations.

### 6.5.1 Automatic Annotation

The recognition of action correctly based on human and object features remains challenging. A successful HOI recognition model must accurately 1) localize and recognize each interacted entity (human, object), and 2) predict the interaction classes (verb). Both are difficult in visual HOI recognition. In our automatic annotation work, we use instance-centric ICAN [130] module which we referred in

Section 6.4.1 to detect the HOI triplet automatically. The module dynamically highlights the regions of object and related attributes. Their approach shows a sizable performance boost compared with other state-of-the-art techniques. It also can detect multiple action of individual and multiple person. The visual examples of detected HOI on lifelog dataset are shown on Figure 6.6.

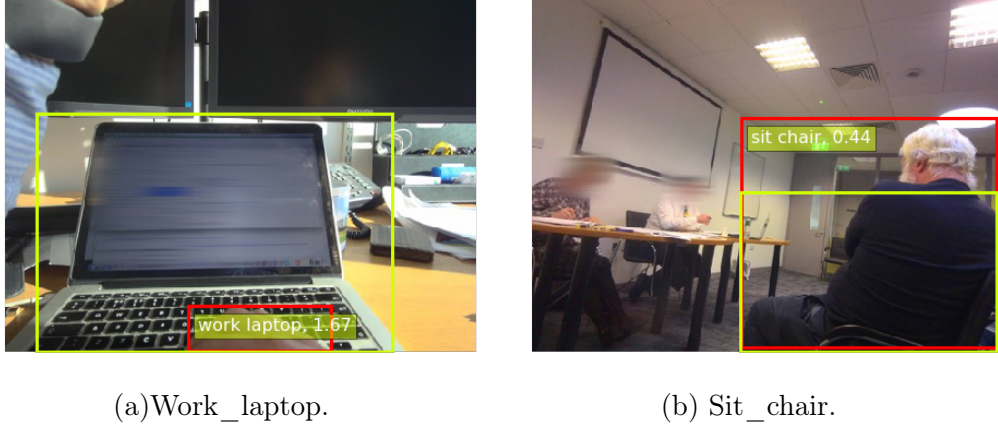


Figure 6.6: Examples of HOI automatic annotation in NTCIR13 lifelog Dataset

### 6.5.2 Manual Annotation

We built an automatic annotation tool for HOI annotation manually. The dataset we used for annotation is NTCIR13-lifelog dataset. We annotated 45 days data based on the prediction  $\langle subject, interaction, object \rangle$ . The annotation work is not as trivial as drawing bounding boxes around all the humans and objects associated with the interaction. So we only annotated the images with human recognized HOI following requirements. There were 8 volunteers arranged to finish annotation of the dataset, which took two weeks. The interface of human object interaction annotation platform(<http://lifelog-ci.computing.dcu.ie/login>) is shown in Figure 6.7 and the examples for HOI annotation are shown in Figure 6.8.

We annotated the human object interaction in three categories, the detailed explanation is listed as follows:

- HOI (Human Object Interaction): Human Object Interaction, “Human” is the

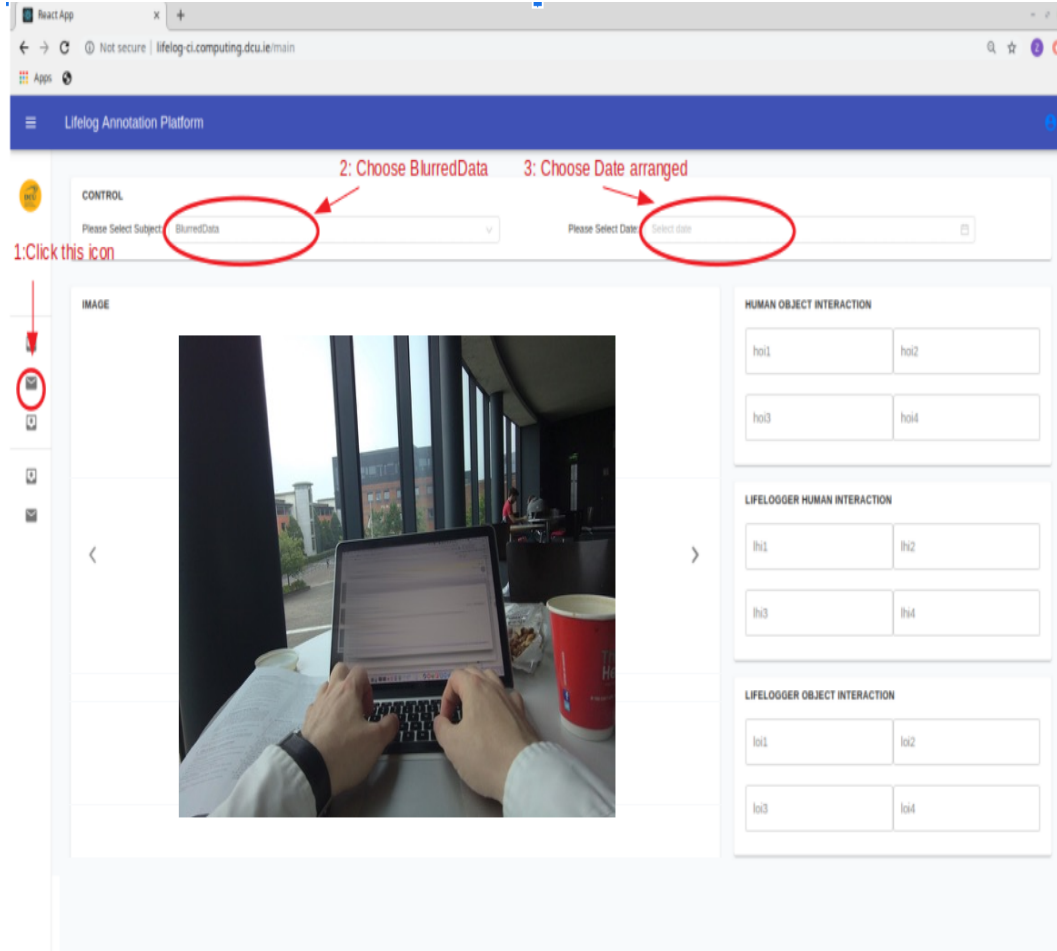


Figure 6.7: HOI annotation tool

others, including man, woman, boy, girl.

- LOI (Lifelogger Object Interaction): It means the interaction between Lifelogger and objects. For example - lifelogger hold apple, lifelogger use laptop.
- LHI (Lifelogger Human Interaction): It means the interaction between Lifelogger and other person. For example - lifelogger talk with man, woman talk with woman.

In order to better view and compare the performance of HOI extraction from these two annotation approaches, the visualization for the distribution of selected HOI based on NTCIR13-lifelog dataset are shown in Figure 6.9 and Figure 6.10.


HOI Annotations	Images
LHI: no LOI: lifelogger hold coffee HOI: no	
LHI: lifelogger talk with man LOI: no HOI: no	
LHI: Lifelogger give book to man LOI: Lifelogger hold book HOI: man hold book	

Figure 6.8: The Example for Manual HOI Annotation

From the count of "work laptop", we realized that the automatic annotation is not so ideal now, because some HOIs from some lifelog images are missed which could be caused by quality of images.

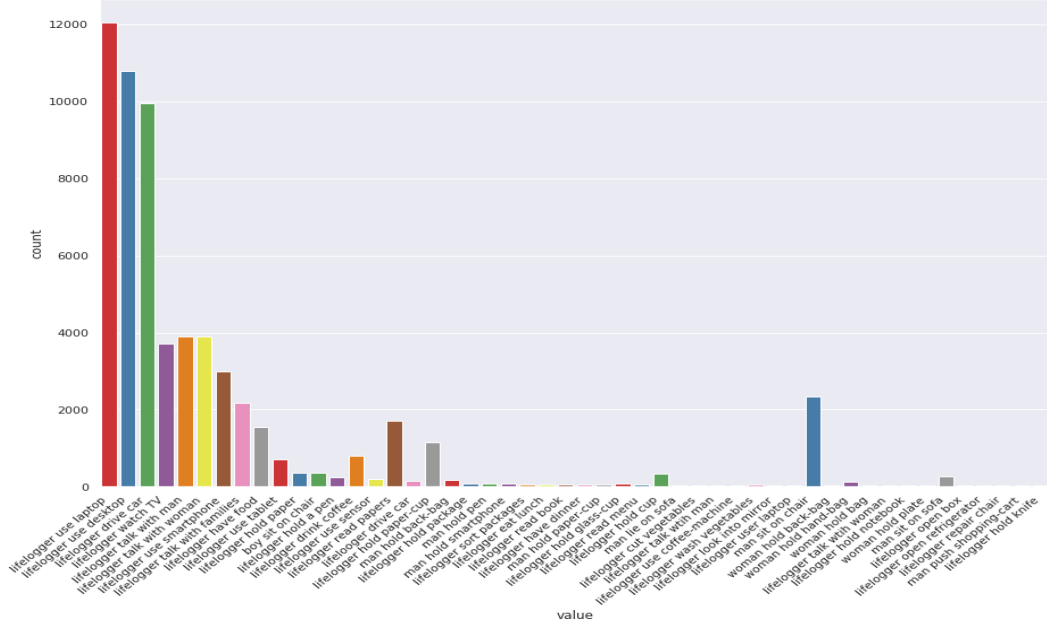


Figure 6.9: HOI Histogram for selected manual HOI annotation

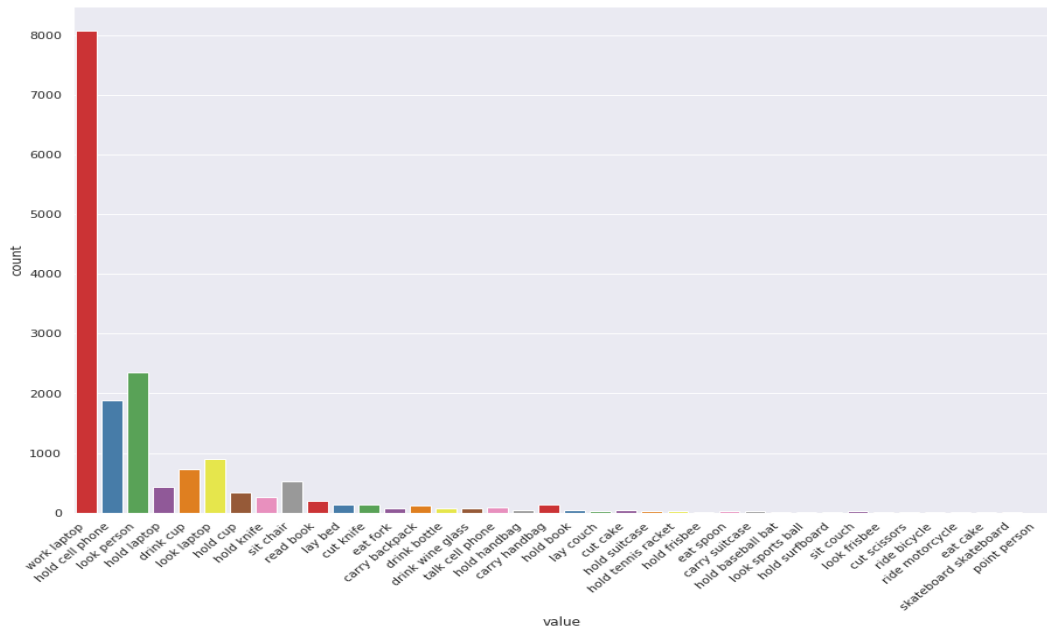


Figure 6.10: HOI Histogram for selected automatic HOI annotation



## 6.6 Experimental Results

We evaluated the performance of our proposed HOI model on HOI-S video dataset and lifelog dataset.

### 6.6.1 Experimental Datasets

We tested the causality enhancement information retrieval in a video dataset and NTCIR13 dataset. We utilised both training and test datasets for this evaluation. Firstly, for training both the iCAN model and image captioning models we used the following two public datasets:

**V-COCO** is a first HOI detection benchmark and dataset that provides detailed HOI annotations. It is a subset of popular MS-COCO [150] and it includes a total of 10,346 images containing 16,199 human instances. Each person is annotated with a binary label vector for 26 different actions. Each person can perform multiple actions at the same time, for example a man is walking and holding a handbag.

**HICO-DET** is a subset of the HICO dataset [167]. It contains 600 HOI categories over 80 object categories and provides more than 150K annotated instances of human-object pairs.

And for evaluating the HOI inference retrieval, we used the following two datasets:

**HOI-S dataset:** This is a dataset that we generated for this work, which was collected by volunteers in Japan using three different surveillance cameras arranged in a formation similar to that shown in Figure 6.2. It includes 3.9G of video data and we annotated these dataset using the Labellmg annotation tool. The dataset shows an instance of a man holding a bag who swaps his bag with the bag of another. In order to evaluate our approaches, we trimmed this video and cut them into 2 test sequences.

**NTCIR13 Lifelog dataset:** This dataset consists of a large collection of

multimodal lifelog data over 90 days which was collected by two lifeloggers.

### 6.6.2 Implementation Details

We evaluate three methods based on our work for the experiments of this proposed approach:

- The first method is using the image captioning model which is trained on GPU Quadro P5000 using CNN and LSTM, as described in this section (referred to as **LSTM+CNN** in the following result tables);
- The second method is built based on iCAN [130] with a feature backbone of ResNet 50 FPN. The main purpose of the modification is generating the HOI output directly from video frames and lifelog captured images and filtering all HOIs to choose the high-ordered human-centric ones which will be used to represent the frames and images. The iCAN model using the V-COCO and HICO-DET datasets and applied the pretrained model to extract the HOI triplets which are used to calculate the causality. Then, we extract the HOI of HOI-S and lifelog data to retrieve the relevant human object interactions;
- The final method is a **Human-in-Loop**, defined as a model with human interaction, which is employed in order to integrate a best-possible HOI detection, to check the captioning result and give relevant feedback which will be used to modify the caption results.

### 6.6.3 Results Comparison Using Lifelog and Video Dataset

In this section, we compared our proposed framework and evaluated our approach based on the retrieval results from the three approaches previously described. As discussed in Section 6.4, the ultimate goal of our model is to infer information transfer between different HOIs in the lifelog and videos dataset by analysing the relationship between objects and humans. Specifically, we compared the approaches (described

in previous section) using the HOI-S and NTCIR-13 lifelog datasets in terms of F1 measure and map@10.

## Video Dataset

The sample of transfer entropy scores calculated based on HOI inference results obtained from the iCAN model is shown in Table 6.2 and the corresponding illustrations of the extracted images are shown in Figure 6.11. Additionally, the evaluation results of retrieval performance on the HOI-S dataset is shown in Table 6.3 for all the three approaches described. As can be seen and expected, the Human-in-Loop approach significantly outperforms the two automated approaches, which highlights the importance of highly effective HOI generation process. Additionally, the iCAN approach clearly outperforms (0.21 compared to 0.14) the LSTM-CNN approach. The reason for this is under investigation, but our conjecture is that the the difference in performance is due to the nature of the videos in the HOI-S dataset not being as suitable for the HOI extraction process. Finally Table 6.4 shows direct comparison between the two automated approaches in terms of F1 measure across HOI-S datasets and gives a clear indication of how much the iCAN approach performs better than the LSTM-CNN approach.

Table 6.2: Example of calculated Transfor Entropy scores on HOI-S dataset

HOI_start	HOI_end	TE_score
Woman 1_carry_handbag	Man 4_carry_backpack	0.063
Woman 1_carry_handbag	Man 1_carry_backpack	0.060
Woman 1_carry_handbag	Man 5_carry_backpack	0.041
Woman 1_carry_handbag	Woman 2_carry_backpack	0.041
Woman 1_carry_handbag	Man 6_carry_handbag	0.024
Woman 1_carry_handbag	Man 6_carry_shoulder-bag	0.018

Table 6.3: Comparison results in HOI-S dataset

Approach	Map@10
LSTM-CNN + Causality	0.14
ICAN + Causality	0.21
Human-in-Loop + Causality	<b>0.8</b>

Table 6.4: Comparison between both automated methods

Approach	F1
LSTM-CNN + Causality	0.13
ICAN + Causality	0.25

### Lifelog Dataset

We examined our method in lifelog moment retrieval task: NTCIR13 LSAT task. We enriched attributes by using the annotated HOIs and re-rank the ranked results based on the Transfer Entropy score. There are 4 approaches employed for the results comparison. The example of queries for LSAT tasks are shown in the Listing 2 in Chapter 3.

- Automatic HOI: this approach uses the ican model to extract HOI.
- Manual HOI: this approach uses manually annotated HOI to enrich the features which can improve the performance of retrieval system.
- Automatic HOI + Causality: this approach uses the ican model to extract HOI and then apply the qualified value of causality from the HOIs , which are extracted from queries and images, into lifelog retrieval.
- Manual HOI + Causality: this approach uses the human annotated HOI and then apply the qualified value of causality from the HOIs , which are extracted from queries and images, into lifelog retrieval.

The evaluation results of retrieval performance on the lifelog dataset is shown in Table 6.6 for all four approaches described. As can be seen and expected, the Human annotated HOI approach significantly outperforms the two automated approaches, which highlights the importance of highly effective HOI generation process. Additionally, the approach with causality score outperforms (0.38 compared to 0.4 on F1 measure) the approach without causality. It is also obvious that the enriched features with HOI improve the accuracy of retrieval.

Table 6.5: Example of calculated Transfor Entropy scores on NTCIR13 lifelog dataset

HOI_start	HOI_end	TE_score
Lifelogger_carry_notebook	Man_carry_notebook	0.056
Lifelogger_carry_notebook	Woman_carry_notebook	0.041

Table 6.6: Overall comparison of different approaches on NTCIR13 dataset

Enriched Attributes & Results			
Approach	MAP@10	CR@10	F1
Baseline	0.45	0.31	0.29
Automatic HOI	0.55	0.42	0.38
Automatic HOI + Causality	0.53	0.38	0.4
Manual HOI	0.8	0.7	0.65
Manual HOI + Causality	0.83	0.75	0.67

#### 6.6.4 Visualization

Here we provide visualizations for some retrieval results as shown in Figure 6.11. The first line shows the keyframes extracted from the video captured by camera2 and the second line shows the video captured by camera3. Red square marked the potential relationship existing between two HOIs.

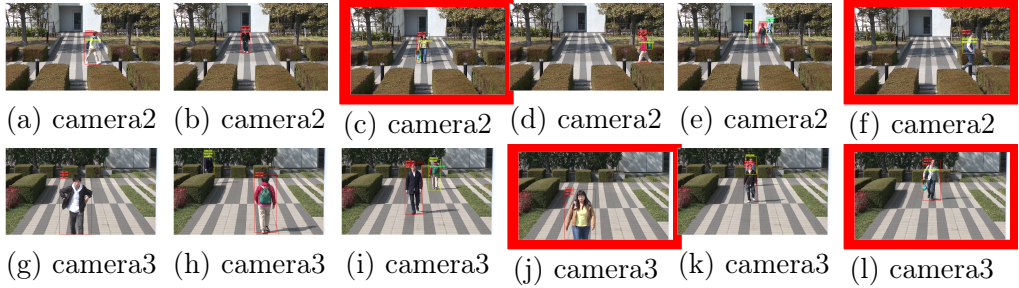


Figure 6.11: Result used the ICAN model and HOI-S video dataset

### 6.7 Discussion & Contribution

In order to answer the RQ3, we proposed a human-centric causality model for HOI retrieval. Our core idea is to retrieve semantically meaningful results from different but relevant camera videos and personal lifelog images dataset. Our work highlights an object instance, which allow us to gather relevant contextual

information facilitating the HOI detection. Our experimental results show that our two automated approaches, the LSTM+CNN and iCAN are significantly outperformed by an approach based on the Human-in-Loop annotation of the HOIs prior to retrieval. This is certainly what one would expect and gives a motivation to further enhance our HOI generation process. The two automated approaches are compared and the iCAN approach performs better than LSTM+CNN, but this in itself is not surprising because the quality of the HOI annotations from the iCAN approach is significantly better than those generated by the LSTM+CNN.

In order to exploit whether HOI recognition and causality in HOI can improve the accuracy of lifelog moment retrieval system, We also applied the HOI recognition and causality inference approaches into LSAT task of NTCIR13-liflog. The results shows the enhanced HOI can advance the retrieval highly but the qualified causality between different HOI enhance the accuracy slightly.

We view this work as an initial implementation and proposal for a method for inferring the relationship (causality) between humans and objects in video and image data. Our proposed approaches evaluated highlight the scope for significant enhancement to retrieval performance for this important and challenging topic. Additionally, this work has focused on a class-agnostic instance-centric approach. We believe that the class-dependent instance-centric attention is a promising future direction, which we will explore, along with the implementation of an end-to-end video retrieval system for various use-cases, such as surveillance video, conventional video retrieval, and semantic retrieval of lifelog data. Finally we intend to integrate transfer entropy calculation into the iCAN model to further enhance the retrieval performance by reducing the noise inherent in the process by reducing the complexity of the approach.

We proposed a novel and practical human-centric causality model for lifelog moment retrieval in this chapter and exploit the performance on large-scale lifelog image retrieval using causality inspired features. Following contributions has been made in our work:

- We proposed and designed a human-centric causality model for HOI retrieval.
- We applied transfer entropy to infer the relationship between different HOIs and create the causality score matrix.
- We exploited the application of HOI recognition and causality between different HOIs in lifelog retrieval systems.

## 6.8 Chapter Summary

After examining these approaches and finding them to be useful, we investigate methods to answer our third research question. We demonstrate that we can address the research question of improving retrieval performance by enriching the standard visual lifelog features by integrating causality-inspired features to provide a state-of-the-art retrieval system. For the question, we proposed a new technique that infers the unseen relationship between two individuals captured by cameras and use it to retrieve relevant video clips or lifelog moment if there is a likely interaction between the two individuals. We begin our research by investigating the existing semantic relation between different human action which can be viewed as causality and exploiting the importance of such semantic relation in video clip and lifelog moment retrieval task. We find that the semantic relation can be qualified and also can be used to enrich the standard visual lifelog features by integrating causality-inspired features to provide a state-of-the-art retrieval system. So we proposed extracting human object interaction and inferring the quantified relationship to enhance the retrieval work. In comparison to previous baseline system of concepts-based lifelog search, we found that integrating casual relations into the content indexing process and using transfer entropy to calculate the causality score leads to improvement in video and lifelog retrieval performance and outperform the previous methods in lifelog retrieval. Even though the approaches using enhanced HOI and causality feature improve the retrieval performance, but sometimes the improvement could be slight when relationship between different

HOI is weak. So in order to boost lifelog retrieval, further we change back to the direction on annotation of applying state-of-the-art text and visual or joint embedding technologies into lifelog query-text matching in next chapter.



## Chapter 7

# Deep Cross-Model Retrieval for Lifelog Retrieval

### 7.1 Introduction

Could text embedding or visual-semantic joint embedding enhance the lifelog retrieval performance? And how much improvement can be achieved beyond the state-of-the-art of lifelog retrieval? In this chapter, we address these questions with two novel approaches based on deep learning architectures. The approaches inventively use pre-trained textual and visual-semantic embeddings for the lifelog retrieval problem and enhance retrieval performance with high-level semantic mapping of user information, which are our main contributions in this chapter.

In Chapter 5, we examined how HOI and causality features are related to each other semantically and how they play a significant role in video and lifelog information retrieval. Enhanced HOI triplet elements have shown benefit for video and lifelog information retrieval, but it ignores the semantic relationship between queries and images, and only focuses on the extracted concepts and combined terms. This is because the method assumes that all categories or tags are known in advance, which does not hold for complex scenes. To improve the retrieval accuracy for realistic and complex scenes, we need to drop the assumption mentioned above and reduce the “semantic gap” between query and image features of lifelog data.

Another issue from Chapter 4 and 5 is the single modalities applied in the retrieval. All visual content must be extracted from images and predicted into

different classes before the ranking part, but in reality the query and the document are in different modality. How can we retrieve images which have the same semantics as the text without modality exchange? And how can we deal with this kinds of cross-modal retrieval? These questions are also addressed in this chapter.

Cross-modal retrieval enables flexible retrieval across different modes, and the key aspect of cross-modal retrieval is measuring similarity between different types of data. Cross-modal retrieval has been proven to be an effective solution when searching over enormous and multi-varied data. Lifelog retrieval is difficult as its data include multiple modes and each modality in itself comprises certain aspects or features. An example of such cross-media retrieval is searching images through a text query. The retrieval system should be able to judiciously determine what modalities (from images, activity, location and biometric data) to use and combine them to make a search more efficient.

In this chapter, we will describe and compare two approaches to address RQ4. The first is lifelog retrieval built on the pre-trained image captioning in Section 7.2. In this caption-based retrieval, we focus on predicting the image captions and matching the query with predicted captions. The second approach is the visual-semantic embedding based lifelog retrieval in Section 7.3. Visual-Semantic embedding mainly aims to find a common latent space where the related visual and textual documents are close to each other, and it is the main task in cross-model retrieval domain [168, 169]. In our visual-semantic retrieval, we employ the deep learning methods for both images and text to match cross-modal documents. In our experiment, both approaches use BERT, which is currently the state-of-the-art word embedding model. As the successful application of transformer attention [170] in NLP, there are a series of approaches proposed for pre-training the generic representation, mainly based on Transformers, Like Bert [171], GPT [172], XLM [173]. But BERT is the most widely used one due to its simplicity and superior performance. So for textual embedding, we choose to use BERT.

## 7.2 Text Embedding for Semantic Retrieval Based on Captions

The first approach is to use pre-trained image captioning. The automatic understanding of semantic correlations between text queries and associated lifelog images as well as their interaction has a great potential for enhanced multimodal lifelog retrieval systems. The semantic correlation represent the actual intention of the user behind the query. When intention cannot be interpreted by the machine, a semantic gap appears. For instance, if the query is 'Nike', the user may be looking for the shoes of 'Nike' brand, but the system cannot capture this intention. Rather, images of 'Nike' brand are returned by the system. To reduce this semantic gap, extracting the semantic features becomes necessary. A typical method resorts to the image captioning, which translates the visual content to text, to bridge the semantic gap between vision and language in lifelog retrieval.

In lifelog retrieval, the searching for images using textual query primarily depends on the metadata (e.g., keywords) of images offered. The similarity between the image and metadata can be calculated and then ranked. This methodology either relies on humans to annotate images or accurate concepts automatically extracted to a great extent. Desired results might not be obtained if there is an error in human annotations of metadata or if the metadata does not define the context behind an image. In our experiment work in last chapter, due to the retrieval work is based on concepts and enhanced Human Object Interactions which may miss the important semantic meaning existing in the images, in order to bridge the semantic gap, we consider to use the image captioning which semantically describes the images in lifelog moment retrieval task at an object level (see 7.1) in this chapter. We will first discuss and compare the state-of-the-art technologies for object instance detection applied in our experiment in next subsection.

### 7.2.1 Object Detection and Attributes Analysis

Extracting objects and corresponding attributes (e.g. blue bag, large room) in images are important in image captioning. In the experiments of Chapter 4 and Chapter 5, the images were annotated with the outputs of a semantic concept detector from Microsoft accessible via their computer vision API<sup>1</sup> which uses Resnet [174] for the object detection and recognition. This has performed well on various object recognition challenges. We reviewed the state-of-the-art bottom-up attention model and Microsoft Computer Vision API, and compared the difference between these two techniques. The detailed comparison in different aspects are listed in Table 7.1. Figure 7.1 shows an example after applying the two detection methods for a random image from our lifelog dataset. When using the bottom-up attention model, the bottom-up region provides the localization of labeled objects from images with complex background. Therefore in this chapter we use the bottom-up attention model which is built upon Faster-RCNN to extract semantic concepts, object labels and object attributes to describe the visual content of lifelog images. We use the pytorch library running on a machine equipped with a Tesla K40m Graphic card and 31 GB RAM for the bottom-up attention model, and it takes about two seconds to extract the visual information from an image.

#### Microsoft Computer Vision API

The Microsoft Computer Vision API is Azure’s computer vision service, which provides developers with access to algorithms that process images and return information based on the visual features the users are interested in. This computer vision service has many different applications, like pedestrian and object detection, adult content filtering, face recognition. Also, this cloud-based Computer Vision API provides developers with access to algorithms for processing images and returning information (e.g. the size of image, extract concepts), by uploading an image or specifying an image URL. Microsoft Computer Vision algorithms can

---

<sup>1</sup><https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/>

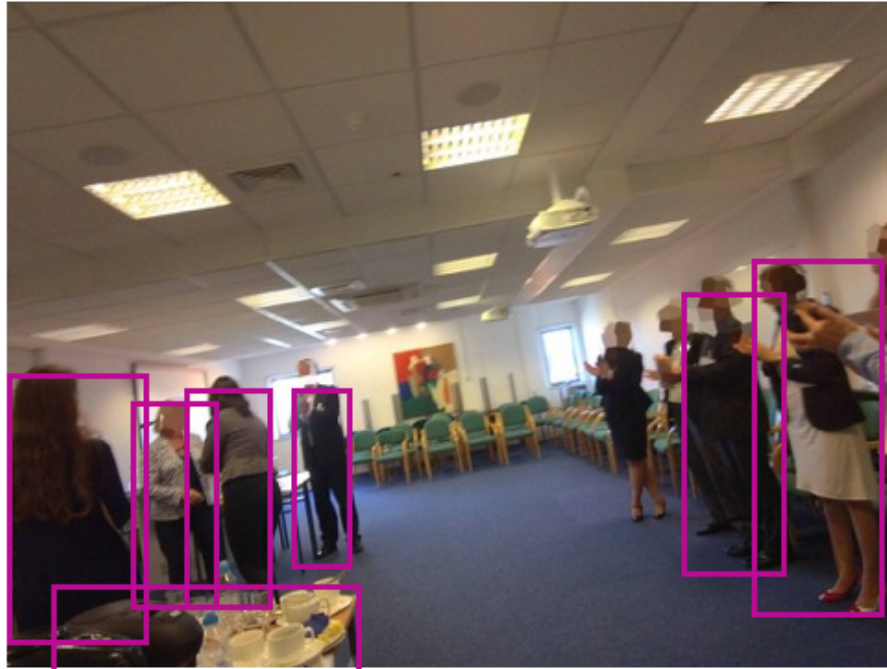
analyze visual content in different ways based on inputs and user choices.

### Bottom-UP Attention Model

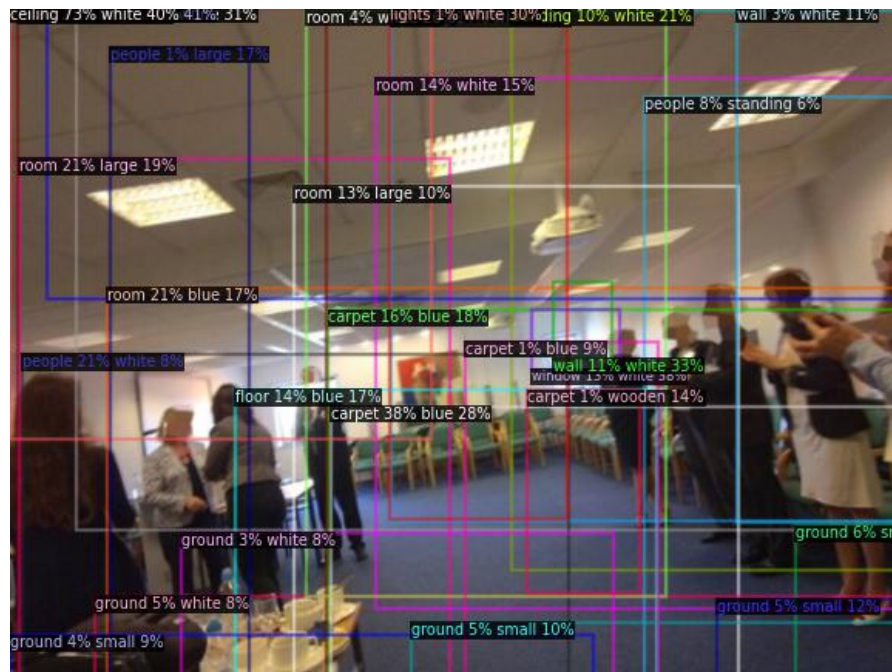
The bottom-up visual attention has been studied in image captioning and visual question answering to deeper understand the fine-grained analysis of semantic reasoning. The bottom-up attention [122] is designed to retrieve a local region information (including spatial information, the position information in a feature map) and global region information (fully connected feature) at the same time. The bottom-up mechanism is based on Faster R-CNN [148]. Each region proposal in Faster R-CNN has an associated feature vector and the corresponding weights are determined by a top-down mechanism. The model generates bounding boxes as regions of objects and feature vectors as object descriptors. In our image captioning, Faster-RCNN is adopted to detect and encode image regions at the object level. Image-text similarity is then obtained by aggregating similarity scores from all word-region pairs. The Figure 7.1 show an example of visual output using Bottom-up Attention model and Microsoft Computer Vision API. From Table 7.1, we know the Microsoft API uses the COCO(201-2017) dataset, which is a large-scale object detection, segmentation, and captioning dataset. It includes 80 objects and 1.5 million object instances. Bottom-up-attention use Visual Genome dataset. It is a dataset which enables the modeling of the interaction and relationships between objects in an image for cognitive tasks.

Table 7.1: Comparison of Object Extraction

Techniques	Bottom-up Attention	Microsoft API
Object Classes	1600	80
Dataset	Genome	COCO(2014-2017)
Size	108K	1.5M
Backbone	Faster R-CNN	Resnet



(a) Meeting Room(Microsoft Computer Vision API).



(a) Meeting Room(Bottom-up Attention model).

Figure 7.1: The Bottom-up Attention model and Microsoft Computer Vision API labeled the image with objects class and corresponding attribute class, and the Microsoft Computer Vision API only focused on label image with object categories.

### 7.2.2 Model Architecture

The main idea of our approach is to efficiently generate image captioning using state-of-the-art techniques and using a BERT-based approach [171] to calculate the similarity over the entire lifelog collection. The overview for image captioning approach is shown in Figure 7.2. The system consists of multiple stages. In the first stage, we apply image captioning generation methods to generate the enhanced captions of each image. Following that, a BERT-based similarity method is used to calculate the similarity between input query and captions extracted from images. In a final step, the retrieved results are ranked based on the similarity.

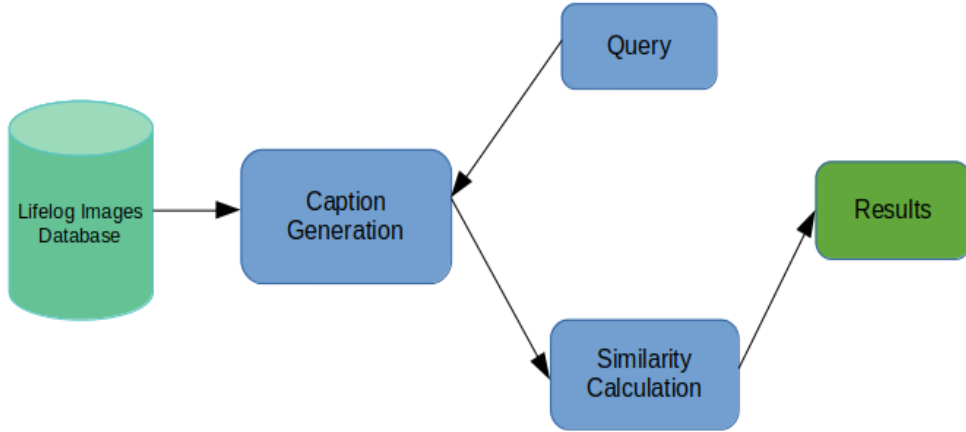


Figure 7.2: Architecture overview of the caption-based search system

### 7.2.3 Updated Image Captioning Generation

The updated image captioning model is transformer-based. The image captioning helps machine understand visual information and express it in natural language, and has attracted increasingly interest in computer vision. Image captioning generally describes the contents based on the semantic relations and relative positions of objects in an image. Predicted captioning needs to describe the implicit attributes of the item which are extracted using bottom-up attention

model and cannot be easily localized by object detectors. Commonly, the approach is to use a nearest neighbor algorithm. This takes two or more vectors, and calculates the distance (or similarity) between them in an efficient manner.

## Feature Extraction

Figure 7.3 shows our bottom-up attention model for feature extraction. A common CNN model architecture has a number of convolution and pooling layers stacked one after the other. The Pooling layers are used to reduce the dimensions of the feature maps and it summarises the features present in a region of the feature map generated by a convolution layer. The object descriptor is taken from the Pooling5 layer of the model and the shape is 2048 dimensions. The number of object descriptors from an image depends on how many objects are detected, which can range from 10 to 100. In addition to the object descriptors, we also include the height, width and coordinates of every bounding box into each feature vector. All these vectors are saved for further processing.

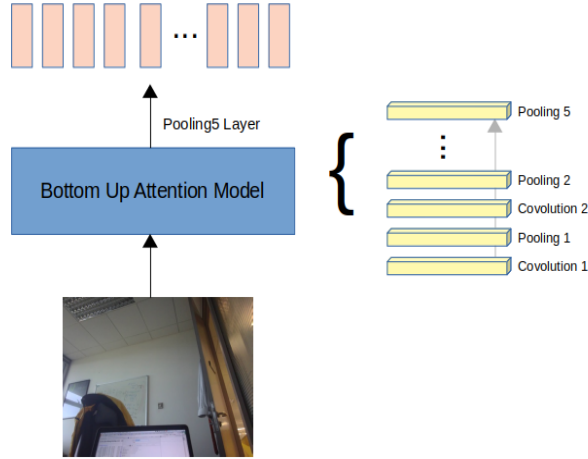


Figure 7.3: Architecture overview of feature extraction



## The M2 Model

The M2 (Meshed-Memory Transformer) model [123], a Transformer-based image captioning model, uses a transformer encoder for self-attention on visual features and a transformer decoder for masked self-attention on caption tokens and encoder-decoder attention. A linear layer projects visual features into the usually lower-dimensional representation space of the encoder. Noticeably, the M2 model can exploits a prior knowledge by incorporating a region encoding approach and a meshed connectivity between encoding and decoding modules. Besides, its experimental results is ranked first among published algorithms in COCO Captions online test server.

## Captioning Prediction

Our captioning generation is based on the M2 model. The input of the M2 model is visual features from images and text features of captions and then output is the probability between the image and caption. We choose the highest probability as our prediction result. The baseline architecture of Captioning Prediction is shown in Table 7.4. Figures 7.5 illustrates the examples of predicted captions using lifelog images. The captions extracted can better describe the content of images as would a human do if caption.

### 7.2.4 Neural Embedding - BERT

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. It is mainly built as part of fine-tuning for the specific classification objective. BERT is basically a trained Transformer Encoder stack and has its own tokenizer, and vocabulary. We use its tokenizer and prepare the documents in a way that BERT expects. The Figure 7.6 shows the main architecture of the BERT Model.

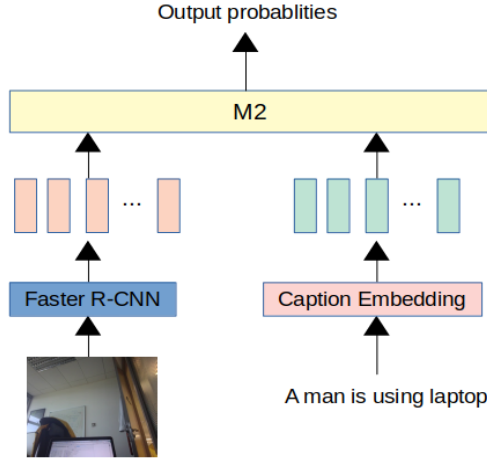


Figure 7.4: Baseline Architecture of Captioning Prediction

### 7.2.5 Semantic Similarity

Semantic similarity is a task to find the similarity between different documents, including type of audio, visual, text data. The most general method to address such semantic search is to map the each encoded document vector into a single vector space and determine which semantically similar sentences are close via calculating the distance between the vectors. After selecting and interactive computation, we choose to use the Sentence-BERT [175] to calculate the semantic similarity. Sentence-BERT a modification of the pretrained BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. The reason we choose Sentence-BERT is its higher efficiency for finding the most similar pair than BERT while maintaining the accuracy of BERT. The overview of the process to calculate the similarity is shown in Figure 7.7. This is accomplished by feeding a pre-trained Sentence-BERT model with sentences and queries while computing similarity scores.

The key is computing the semantic relevance between a given natural-language query and a specific lifelog image. In our experiment, the cosine-similarity distance is employed to calculate the similarity between the captioning result and text query,

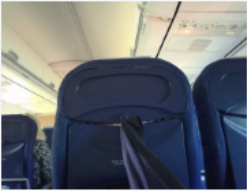
ImageID	Image	Caption
20160827_104351_000		an airplane room with lots of blue seats.
20160930_131650_000		a collage of pictures of a man in a collage.
20160917_181756_000		a collage of pictures of cars in the sky.
20160825_075658_000		a group of computer monitors sitting on a desk.
20160925_174256_000		a collage of pictures of the inside of a garage.
20160921_135128_000		a view of a street with a cloudy sky.

Figure 7.5: Example of updated captioning results

and thus semantically similar sentences can be found. These similarity measures can be performed extremely efficiently on semantic similarity search as well as for clustering. Given an captions and query denoted as  $Q$  and  $C$ , and fed them into the Sentence-BERT model. The cosine function are used to calculate the relevance score. The formula are shown below:

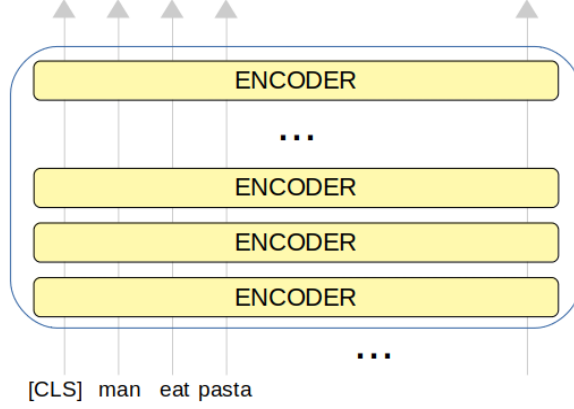


Figure 7.6: The architecture of BERT. [CLS] is the first token of every sequence.

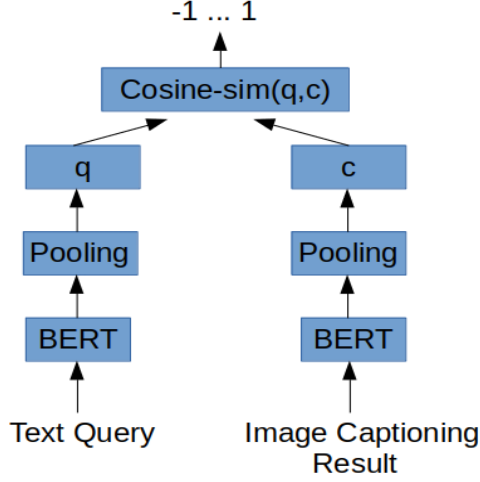


Figure 7.7: The calculation of similarity

$$\cos(\mathbf{q}, \mathbf{c}) = \frac{\mathbf{q}\mathbf{c}}{\|\mathbf{q}\|\|\mathbf{c}\|} = \frac{\sum_{i=1}^n \mathbf{q}_i \mathbf{c}_i}{\sqrt{\sum_{i=1}^n (\mathbf{q}_i)^2} \sqrt{\sum_{i=1}^n (\mathbf{c}_i)^2}} \quad (7.1)$$

Table 7.2 show the an example for the similarity between different sentences. The cosine similarity score between query and captions predicted using the pre-trained model are calculated.

Table 7.2: Example of the similarity calculation between query and image captions. The query: A man is eating pasta.

The top 5 most similar sentences in corpus	Score
A man is eating food	0.5777
A man is eating a piece of bread	0.4986
A man is driving a car	0.1581
A man is riding a white horse on enclosed ground	0.1474
Two men are pusing carts through woods	0.0892

### 7.3 Joint Embedding for Cross-Modal Retrieval in Lifelog Retrieval

The second approach is the visual-semantic embedding based lifelog retrieval. The purpose of LMRT (ImageCLEF) and LSAT (NTCIR) task described in Section 2.3.2 is matching text queries to images and there exists a “media gap” and “semantic gap” which means the representations of different media type are inconsistent in type and semantics. The main challenge in this image-text retrieval task is matching data from different modalities semantically. The cross-modal retrieval is designed for such scenarios where the queries and retrieval results are from different media types. Currently most recent image-text retrieval methods offer great promise by learning deep representations aligned across modalities. In our work, we follow the state-of-the-art techniques to learn robust Visual-Semantic joint representation which is used to match the text query to lifelog images. There are two types of queries we must consider: how to learn two projection matrices to map multimodal data into a common feature space, in which cross-modal data matching can be performed, how to design a deep neural network that can project the query into a visual feature space by vectoring the query. In our experiment, the text query and image are encoded as text and visual embeddings using different encoders firstly and then pass them to a learned common latent space to calculate the relevance. The overview of the steps are shown in Figure 7.8.

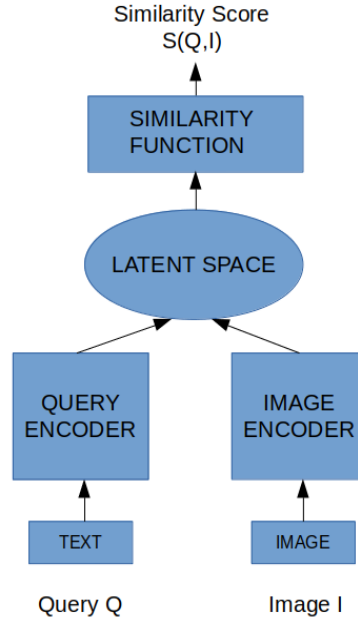


Figure 7.8: Overview of Matching Procedure

### 7.3.1 Visual Representation

Given an lifelog images, how to encode the image to the representation is a first task we should consider. Deep Neural Networks (DNNs) is a powerful technique for learning representations from data, in a hierarchical and compositional way. It has been successfully applied in several fields such as image classification, speech recognition, etc. In this work, we choose to utilise DNNs model to generate the image region and the corresponding representation.

#### Resnet152

In our experiment, we choose to use Resnet152 [174] to extract the visual features. Resnet is belongs to DNN and was proposed by researchers at Microsoft Research. It introduced a new architecture called Residual Network is on the most powerful deep neural network which has achieved excellent performance results in computer vision domain. Resnet152 is trained on the ImageNet dataset and the residual nets with a depth of up to 152 layers. The reason why we select the CNN technique of Resnet152 is because it achieves the best accuracy among Resnet family members. Figure 7.9

illustrates the transfer learning structure for a single resnet152 network. In this Figure, the pretrained resnet152 networks perform as feature extractors for generic image features and the two last layers are fully-connected layers for classification. In our work, we applied the resnet152 to extract the visual representation.



Figure 7.9: The transfer learning structure for a single CNN.

For the resnet152 model, we take the input of the classification layer as the feature, which has a dimensionality of 2,048. For a given lifelog image, it is resized to  $256 \times 256$  size and then fed it into the resnet152 network to extract last hidden layer of which a shape of  $2,048 \times 1$ .

### 7.3.2 Text-query Representation

When classification is the objective, BoW with tf-idf weighted one-hot word vectors and traditional approaches such as SVM should be the first thing to try. It will establish a baseline we can aim to beat with newer approaches such as BERT. BERT yields high quality results at some expense. But faster and lighter versions of BERT are being explored constantly, and compute is getting cheaper as well with cloud options. Plus BERT embeddings are not limited to producing a sentence vector for classification and the one-hot and fast text embedding don't work as well as on BERT for those other use cases. So in our experiment, we still use the Sentence-BERT encoder which is used to encode an input sentence as a feature vector.

### 7.3.3 Unsupervised Cross-modal Common Vector Space Learning

The main objective of the system is to transform the two different features (image and text, separately) to other features laying in a joint space and compare the

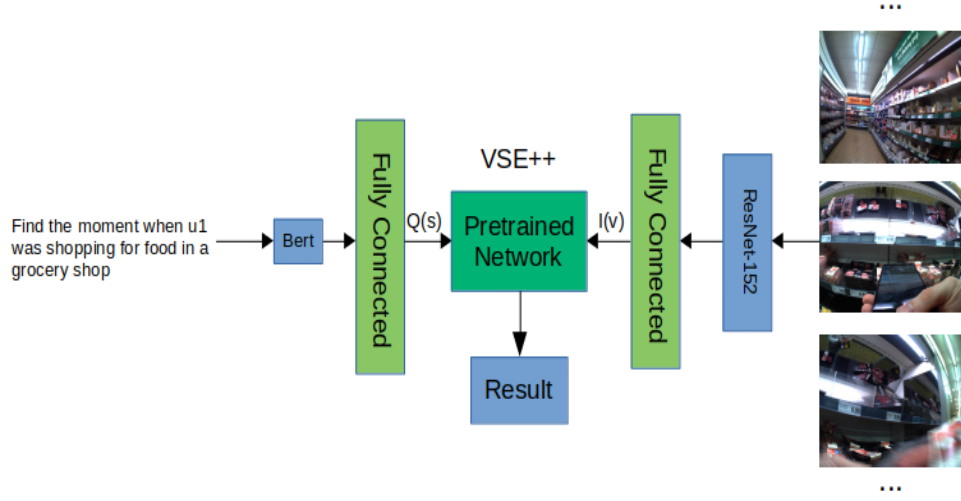


Figure 7.10: The left to VSE++ is using BERT to extract text embedding. The right to VSE++ is applying resnet152 to extract 2048 dimensions visual features. The results is calculated using the cosine-similarity

similarity between these two features. After the text and visual embedding are extracted, how to match the visual and text embedding is the third point we should consider. As we describe mentioned before, image features and the text features could be transformed to the same joint features, in the same space and then calculate the similarity. Common vector space learning based on methods are currently the mainstream in cross-modal retrieval. Among many choice of common vector space learning algorithm, the (Visual-Semantic Embedding)VSE++ [176] has shown impressive performance in image classification and text-based image retrieval tasks. The VSE++ models explicitly maps images into a rich semantic space, with the goal that images of the same category are mapped to nearby location and text descriptions are embedded in such a common vector space with respect to some measure of similarity. The VSE++ model has trained on MS-COCO and Flickr30K dataset and performed well in mapping images to text descriptions. The prototype of this assumption is shown in 7.10. Given a text query expressed by a natural-language sentences of  $Q\{w_1, w_2, \dots, w_l\}$ , we aim to match the sentence and retrieve lifelog images relevant with respect to the query from a collection of  $n$  unlabeled images  $I\{i_1, i_2, \dots, i_n\}$ .



We focus on query representation learning that predicts  $s$  from the query. Meanwhile,  $v$  can be instantiated using either deep CNN features or concept vectors as exploited in previous works.

## 7.4 Experiment Setting

There are two approaches we examined, so two settings are listed, one for lifelog retrieval based on only text embedding and another one based on visual-semantic joint embedding. It is the first time these techniques have been applied to lifelog data.

### 7.4.1 Setting 1 - Text Embedding

We use the following step to do the experiments:

- Visual embedding: Bottom-up Attention model is used to extract visual features with attributes from a given image.
- Captions and query embedding: We choose M2-based image captioning model [123] to predict the corresponding captions.
- Matching: Sentence-BERT [175] vectorized the sentences and captions to calculate and rank the similarity between them.

### 7.4.2 Setting 2 - VSE++

We used the implementation of VSE++ [176] model to match the query and images. We used Sentence-BERT for feature extraction from query sentences and the ResNet-152 models for feature extraction from images.

- Visual Embedding: Resnet152 predicts the CNN feature of ResNet-152 for a given image.
- Query Embedding: Sentence-BERT decodes the text feature to text embedding which will be used to calculate the similarity in next step.

- **Matching:** Project the visual and query embedding into the learned latent space to match them.

## 7.5 Experiments Results

In order to test the approaches of the semantic retrieval based on the image captioning model and cross-modal retrieval model, we applied this method on NTCIR13-lifelog LSAT task (Dataset: 90 days, 114547 images; Topics: 23 topics). Also, we evaluate the performance improvement of our proposed approach by comparing the search results to the baseline approach from Chapter 4. We list the key features for all approaches as follows:

- **Baseline:** For the baseline tool we choose LIFER [146] system which is a concept-based filter using semantic concepts. These semantic concepts are automatically extracted from lifelog images content to represent their visual information. Users are able to filter images content using a single concept or the combination of concepts. This tool could help users to speed up the validation process by providing a list of related images candidates. But this tool suffers the semantic gap problem, which is to successfully describe visual content from low-level feature analysis to semantic content representation of images.
- **Enriched Attributes:** We choose the approach (Automatic HOI + Causality) applied in Chapter 5: this approach is using the human annotated HOI and also applied the qualified value of causality between the HOI in queries and images.
- **Caption:** This is first approach which adapt the image captioning model to covert lifelog into text description and then using sentence-BERT to match the queries and images. This approach is described in Section 7.2.
- **Joint Embedding:** This is second approach which adapt the visual-semantic

model to match the encoded text query and visual images and then project them in the common latent space to calculate the cross-modal similarity which is obtained from the cosine distances. This approach is described in Section 7.3.

## Results Comparison

The evaluation results from two approaches described above are shown in Table 7.3. As can be seen, the approach based on joint embedding HOI approach significantly outperforms (0.41 compared to 0.57 on F1 measure) the one based on captions, which highlight the importance of highly effective learned cross-modal embedding. Additionally, compared with the baseline and enriched attribute approach (from Chapter 5), it is clearly outperforms both of them. There is a classification step to turn the image into captions in the first approach, which could cause the retrieval accuracy to be lower than second approach. Table 7.4 shows the comparison of two approaches on P@10 score. As all LSAT lifelog queries are finding the moment of lifelogger’s behavior, the ground-truth of the these queries are top 10 images. So the cut-off at 10 will be better to compare the performance of our two approaches. Besides, we found that the captioning-based approach could accurately detect specific concepts and scene recognition in a query sentence. On the other hand, in the visual-semantic embedding approach, we found that phrases including verbs, prepositions, and relationship between two objects (people and objects) were captured relatively well.

Table 7.3: Overall comparison of different approaches on NTCIR13 dataset

Lifelog Moment Retrieval			
Approach	MAP@10	CR@10	F1
Baseline(interactive)	0.37	0.3	0.29
Enriched Attributes (automatic)	0.55	0.42	0.38
Caption	0.6	0.45	0.41
Joint Embedding	0.71	0.66	0.57

Table 7.4: The P@10 of 10 query sentences evaluated in the LSAT sub-task NTCIR13-Lifelog and comparison of video retrieval performance (average precision) between caption-based and visual-semantic joint embedding approaches

Query-id	Description	Caption	Joint Embedding
001	Find the moments when I was eating lunch.	0.65	0.88
002	Find moments when I was gardening in my home	0.12	0.23
003	Find the moment when I was visiting a castle at night	0.51	0.67
004	Find the moments when I was drinking coffee in a cafe	0.6	0.7
005	Find the moments when I was outside at sunset	0.56	0.64
006	Find the moments when I visited a graveyard	0.54	0.43
007	Find the moments when I was lecturing to a group of people in a classroom environment	0.35	0.55
008	Find all the moments when I was grocery shopping	0.62	0.68
009	Find the moments when I worked at home late at night	0.67	0.71
010	Find the moments when I was working on the computer at my office desk	0.57	0.85

### 7.5.1 Discussion & Contribution

In this chapter, we proposed two novel approaches of applying the state-of-the-art caption extraction model and visual-semantic embedding model into lifelog retrieval systems. We evaluated the effectiveness of the two novel approaches and also discussed the performance improvement for the enhanced retrieval system. As discussed in section 7.1, this work aims to improve the quality of feature matching and to reduce the "semantic gap" between text query and images. We compared our proposed text and visual-semantic embedding model approaches in performing lifelog moment retrieval task between text query and images. We also compared the results with our previous baseline search engine and causality-based retrieval approach. We hoped that these two approach would out-perform the baseline and

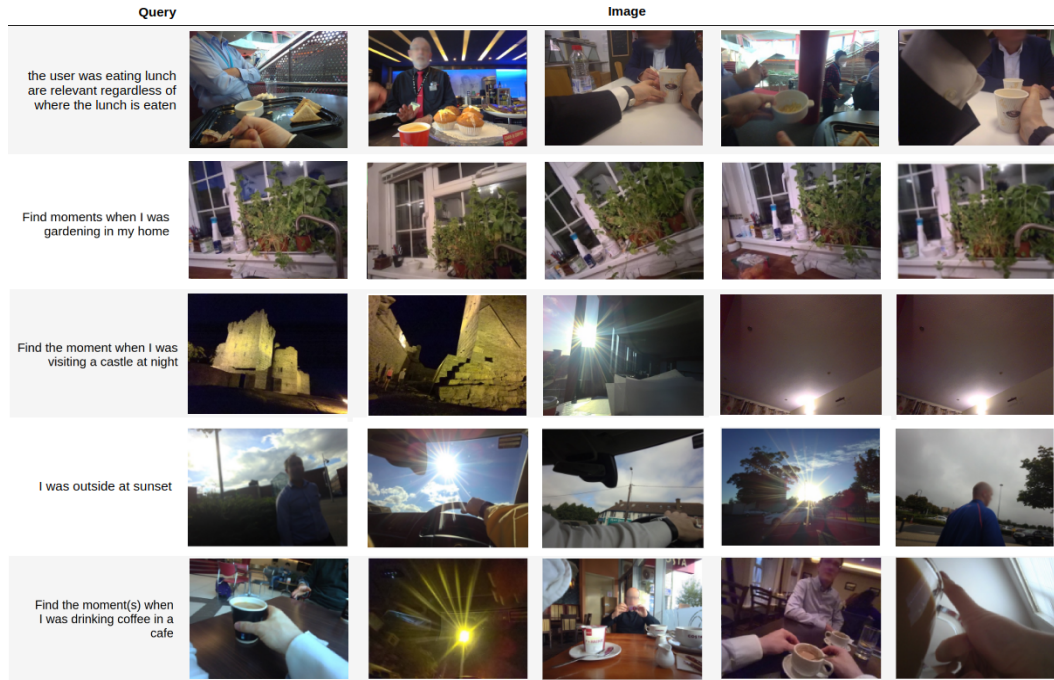


Figure 7.11: Selected example of Top 5 Retrieved Results on caption based retrieval (text embedding)

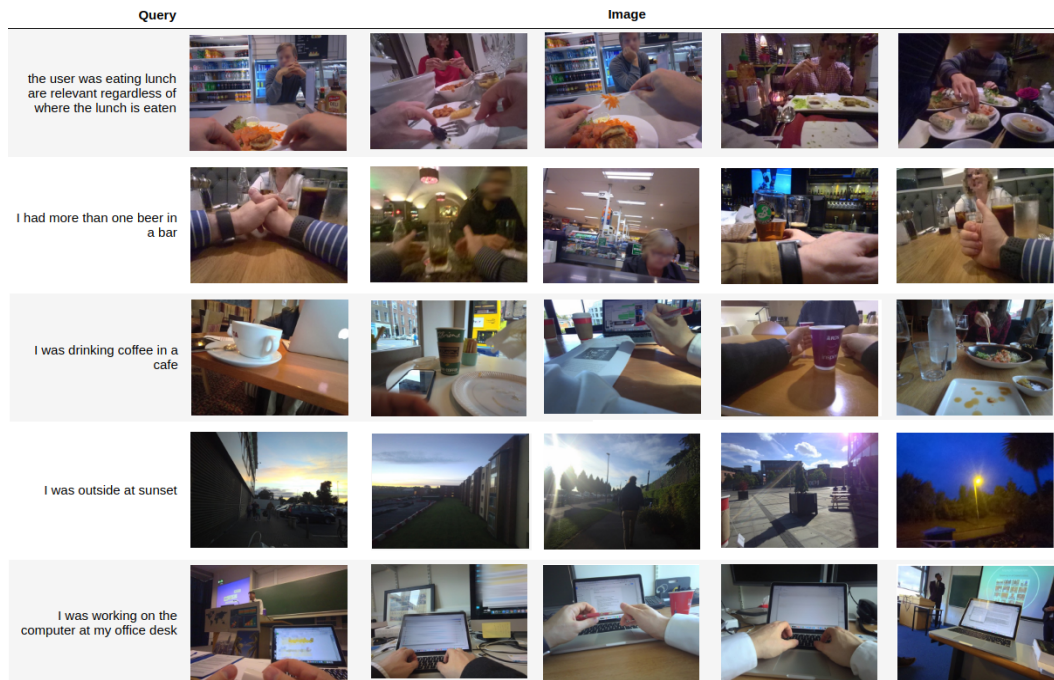


Figure 7.12: Selected example of Top 5 Retrieved Result using the Cross-Modal Based Retrieval (visual-semantic embedding)

causality-based retrieval in terms of effectiveness which could be used to improve the retrieval performance. Figure 7.11 illustrates the selected results using the Image Captioning approach and Figure 7.12 illustrate the selected results using the cross-modal retrieval. Obviously, the approach using visual-semantic joint embedding can retrieve accurately results than caption-based approach. In this figure, the text query is shown on the left side and the retrieved images are shown on the right side. The comparison of the top 5 results retrieved using two approaches clearly demonstrates that our proposed semantic-visual embedding approaches can increase the performance of lifelog moment retrieval. As shown in Table 7.3, with the support of the embedding techniques, the performance of lifelog moment retrieval is improved by 19% than causality enriched attributes (automatic). Therefore we can conclude that our approach could help to reduce "semantic gap" between text query and images, and also boost the retrieval performance because it provides more reliable feature matching after applying the embedding model on lifelog moment retrieval task.

Besides, a user-friendly interface for the retrieval based on these two approaches are not available and this will be our future work. Even though the retrieval results are improved by the two approaches proposed, there are still some drawbacks existing in these approaches:

- The models we employed are trained on public datasets, which are different to lifelog datasets (lifelog dataset is from egocentric vision, it infers the activities from liflogger and others). Currently a large number of lifelog images which can support visual-semantic latent space model training has been collected but they lack sufficient annotation.
- After we examined the result, we found if the query infers more detailed and accurate visual descriptions, the accuracy will be higher than concept and HOI based retrieval systems. But if exact locations and time-stamp are given, the interactive system could be better. Therefore it is expected that learning

a joint image-text embedding with spatial-aware, location-aware, temporal-aware could give us better performance.

We address a novel and practical method for lifelog moment retrieval in this chapter and exploit the performance on large-scale lifelog image retrieval using learned caption or visual-semantic embedding. Following contributions has been made in our work:

- We proposed an approach utilizing trained image captioning model and state-of-the-art NLP technique to learn and match the lifelog query and images.
- we applied the joint representations model to learn the visual-text matching and retrieving the queried related images.

## 7.6 Deployment Proposal

In this chapter, the proposed approaches are only query-text matching, but not fully deployed a system. In order to expand these methods into the system we will integrate these approaches into an interactive system as our future work. The system should be comprised of three primary components: a dataset, a retrieval engine and an interactive browser-based user interface. Here described the main work the three components:

- Dataset Layer: Store all the dataset(i.e. metadata, images, human activities) collected. We have such dataset.
- Retrieval Engine: using our proposed method to generate features from the raw and match, rank retrieved results.
- Interface: visualization the retrieval results via an application mechanism.

In order to design the lifelog retrieval system later, there are five different groups of query criteria that should be followed:

- Temporal criteria: a user can specify the date and time, time span, or period (morning, afternoon, etc);
- Scene criteria: a user can specify a query on scene categories (hotel, restaurant, lobby, etc);
- Entity and Action criteria: a user can specify a query on the existence of entities, or actions/activities;
- HOI criteria: a user can specify a query on the triplet of human object interaction;
- Extra criteria: a user can define a query on biometrics data, computer usage information, etc.

All search features retrieve data from the metadata, indices, and the original lifelog data. We also should facilitate the state-of-the-art relevance feedback mechanism so that the user can further provides new search command to refine the results.

## 7.7 Chapter Summary

In this chapter, we exploit whether text embedding or cross-modal joint embedding enhance the retrieval performance. For image-captioning based retrieval, we consider the image captioning first by deep neural network and then matching the captioning with the query and get the ranked results. Our experiments show better results than our previous methods. For cross-modal retrieval, we consider the text query image retrieval using deep neural networks. Different from the method in last two chapters, the methods feed the visual and text embedding into a joint optimization space and compute cross-modality similarity for lifelog retrieval.



## Chapter 8

# Conclusion

### 8.1 Thesis Overview

This thesis applied cutting-edge deep learning techniques in lifelog retrieval. As part of the research, this thesis takes the challenge of a text-image retrieval task. Existing approaches have low accuracy in lifelog moment retrieval due to lack of universal labelled training data. In this thesis, we focus on applying different object detection and image classification approaches for lifelog retrieval. The new methods proposed in this thesis present better retrieval performance than previously proposed approaches. The work covered in this thesis can be summarised as follows:

**Chapter 1** introduced the concept of *lifelog* and the challenges existing in Information Retrieval from Lifelogs. The research questions, hypothesis and objectives of this thesis are presented in this chapter.

**Chapter 2** reviewed conventional information retrieval methodologies and lifelog retrieval systems, detailed the lifelog disciplines and its application. Additionally, the existing approaches of lifelog information retrieval are reviewed and the difference between these approaches with our work are also highlighted in this chapter. According to the literature, we know there are three important parts affect the performance of interactive lifelog retrieval are identified: instances/concepts detection, semantic query-document matching method, and query analysis. These parts form the direction of our research.

**Chapter 3** presented research methodology and answer the four research questions. Additionally, the evaluation methods employed in this research are also

covered in this chapter.

**Chapter 4** presented what is lifelog data and construct three lifelog dataset based on the process including data collection, data organising and data publishing. The organized workshops are listed and the examples of corresponding text query are shown in this chapter.

**Chapter 5** presented the improved baseline search engine that is proposed for lifelogs. The main idea behind this baseline search engine is to provide a starting point for researchers in the area, as well as a documented system for comparative analysis. Besides, in order to follow the steps in action research, we continued to improve the baseline using state-of-the-art techniques and took part in many benchmarking including NTCIR13, Imageclef 2017, Imageclef2018, Imageclef 2019 lifelog tasks. This lifelog retrieval system was obviously improved through these 4 cycles over 3 years using the different test collections.

**Chapter 6** focused on automatically quantifying causality relationships between different HOIs extracted from lifelog image data and its application in lifelog moment retrieval tasks. Some queries with existing the potential semantic and causality relationship can be easily expressed purely by some keywords. So we introduced a Human-Object Interaction (HOI) model to integrate the causal relationship between human and objects to improve the retrieval results.

**Chapter 7** presented and compared two new approaches to automatically match text queries and lifelog images directly. The first is lifelog retrieval built on the pre-trained image captioning based retrieval. In the caption-based retrieval, we focus on predicting the image captions and matching the query with predicted captions. The second approach is the visual-semantic embedding based lifelog retrieval. Visual-semantic embedding mainly aims to find a common latent space where the related visual and textual documents are close to each other, and it is the main task in cross-model retrieval.

The main question here is how to apply semantic retrieval to query images that display realistic and complex scenes, where we cannot assume all the object categories

are known in advance. Besides, our previous approaches belong to single model retrieval domain. All visual content from images must be extracted and labelled with different classes automatically in advance. In reality, queries and documents are different modality. Consequently, the research question becomes how we can retrieve images which have the same semantics as the text without modality exchange? How to deal with this kinds of cross-model retrieval without model exchange will be another big challenge in our lifelog information.

## 8.2 Research Questions and Proposed Solutions

This thesis is mainly focused on exploring personal lifelogging known-item retrieval and applying state-of-the-art techniques to make lifelog more accessible through data collection, organization and visualization. The hypothesis and research questions are revisited, and also examined with respect to the proposed solutions and the experimental results.

### Hypothesis

The application of modern multimodal information retrieval and annotation techniques enhance the performance of lifelog retrieval in a valid repeatable experimental setting.

From the hypothesis, the following research questions are outlined:

- **Research Question 1 (RQ1). Data.**

*How to construct a test collection of lifelog data to support retrieval system comparison and how to create this test collection in a privacy-aware and legally-compliant manner?* In order to construct a valid dataset, such data is unlikely to be available from online sources and needs input multiple willing individuals to share both data and information needs with the community.

For this question, this thesis defines what lifelog data is and proposes multiple principles for building a lifelog dataset. This is the first time the problem of building such a dataset has been deeply discussed by the time of this thesis.

The dataset construct is one of the biggest challenges in lifelog research. The more detailed content are stated in Chapter 4.

- **Research Question 2 (RQ2). Baseline.**

*How to apply traditional information retrieval techniques to multi-modal lifelog data to develop a high-quality baseline search engine?*

For this question, this thesis covers the detailed process of developing baseline search engine for personal life archives, aiming to make personal life archive more searchable and organizable. The preliminary results of comparing this approach with other approaches illustrate the feasibility of the baseline search engine as a tool for getting insights from personal life archives. More detailed information is described in Chapter 5.

- **Research Question 3 (RQ3). Ranked Model based on Enriched HOI(Human Object Interaction) Attributes and causality with different HOIs.**

*Can we improve retrieval performance by enriching the standard visual lifelog features using causality-inspired features to provide a state-of-the-art retrieval system?*

For this question, this thesis proposed a HOI causality model and applied it to enhance result ranking mentioned in Chapter 6 using semantic relations existing in different human behavior. It is called *Causality*. For research question 3, this thesis covers the details for how to extract human object interaction and infer the quantified the relationship to enhance retrieval work.

- **Research Question 4 (RQ4). Embedding Retrieval in Lifelog Retrieval**

*How would text embedding or cross-model joint embedding enhance the retrieval performance?* The performance of lifelog retrieval system can be enhanced by supporting higher-level semantic mapping of user information need into an embedding representation. For this question, we adopted two

methods, one is using BERT to rank lifelog archives. The other one is learning joint representations and predicting retrieval results based on queries. More detailed information are described in Chapter 7.

At this point, we can consider that our hypothesis holds valid because we have answered all research questions positively.

### 8.3 Limitations

Throughout this work, we have attempted to highlight any limitations to our research where relevant. As we have already discussed, the primary limitations have centred around the test collection. This work only uses the NTCIR-lifelog datasets from 1-3 people and therefore we cannot say with certainty that our findings will be equally valid when transferred to a real-world application. However, we also note that it is standard practice within the research community to evaluate lifelog retrieval using this experimentation method, so despite this limitation, we have still engaged in evaluations using best practice within the IR domain. Another limitation of this work is the emphasis on retrieval focused lifelog baseline. Though we selected LIFER system as baseline, state-of-the-art techniques for information retrieval are always changing, so we have keep updating our lifelog baseline.

### 8.4 Future Work

There are several potential directions for future work and further development of the research in this thesis. Some of the interesting extensions and prospects are outlined as following:

#### 8.4.1 Benchmarking

- Novel Datasets. There clearly is an interest in the community to develop retrieval and analytics tools for lifelog datasets, so there is significant potential

for others in the community to further define and release novel datasets of human life-experience data.

- Richer metadata. NTCIR-lifelog dataset released in our previous work is a starting point. Future dataset should be richer. Repeatedly, we have seen that the best performing lifelog retrieval systems enhanced the provided metadata by relying on additional visual concept detectors, or seeking additional sources of metadata to enhance the retrieval performance. We would like to see new approaches to be developed for the creation of semantically rich metadata for multimodal lifelogs, in order to facilitate more effective retrieval.
- Bridge the lexical gap. Many participants found that there was a lexical gap between the terms used by the lifeloggers in their topic descriptions, and the indexed textual content and annotations. This suggests a need for term or query expansion, and the current consideration is that this could be achieved using approaches such as conventional query expansion or word embedding.
- Integrate external WWW contents. This has been used by some participants with positive results. The external content helps to enhance the quality of content annotations or can be used as a form of query enhancement.
- Train our lifelog vision-language dataset which can be used for training the visual-semantic embedding. Currently more and more research focus on the work of training vision-language pre-trained model [177, 178]. The text-images retrieval based on such pre-trained model has become a trend in the future.
- Deploy the proposed system in the LSC'21 (Lifelog Search Challenge in ICMR conference), the fourth lifelog benchmarking challenge, to compare to participants with their interactive lifelog retrieval systems.

### 8.4.2 Retrieval System

#### Advanced Feature Representation

For future developments of the system, there should be a focus to train a lifelog Spatial-Temporal network model based on self-attention mechanism so that the system could learn from labels and lifelog images description based on annotated lifelog data. The self-attention, also known as intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence and it is common to use it to capture temporal information in NLP. However, the premise is that we must annotation the lifelog dataset to built the Image-Text dataset.

#### Advanced Learning to rank

The rank function is the core of information search engine, which provides ability to rank candidate documents based on their relevance to the query. Learning to rank (LTR), a relatively new field, aims to tune more advanced rank functions for more accurate search experience by employing machine learning algorithms. LTR learns to directly rank items by learn a scoring function to predict the probability of a certain item ranking over another item.

LTR models has been widely utilized in search and recommendation systems. However, in many practical scenarios like web search, recommending systems,etc. ranking is typically performed using LTR models operating on a large set of features of which text-based features are but a small subset. In addition, LTR model require training the rank function on-the-fly for each query, which could be very inefficient and computationally expensive.

#### HOI Causality

The implementation and proposal for inferring the relationship between humans and objects in video and image data enhanced the retrieval performance by reducing the

noise inherent in the process. We believe that the class-dependent instance-centric attention, which our approach focused on, is a promising future direction. The work we finished are not enough and we will explore more, along with the implementation of an end-to-end video retrieval system for various use-cases, such as surveillance video, conventional video retrieval, and semantic retrieval of lifelog data.



## Chapter 9

# List of Publications

- [1] Ninh, Van-Tu, Tu-Khiem Le, **Liting Zhou**, Luca Piras, Michael Riegler, P. l Halvorsen, M. T. Tran, M. Lux, Cathal Gurrin, and Duc-Tien Dang-Nguyen. "Overview of ImageCLEF lifelog 2020: lifelog moment retrieval and sport performance lifelog." In CLEF2020 Working Notes, ser. CEUR Workshop Proceedings. Thessaloniki, Greece: CEURWS. Org(<http://ceur-ws.org/>) 2020.
- [2] Gurrin, Cathal, Hideo Joho, Frank Hopfgartner, **Liting Zhou**, Rami Albatal, Graham Healy, and Duc-Tien Dang Nguyen. "Experiments in Lifelog Organisation and Retrieval at NTCIR." In Evaluating Information Retrieval and Access Tasks, pp. 187-203. Springer, Singapore, 2020.
- [3] Le, Tu-Khiem, Van-Tu Ninh, Minh-Triet Tran, Thanh-An Nguyen, Hai-Dang Nguyen, **Liting Zhou**, Graham Healy, and Cathal Gurrin. "LifeSeeker 2.0: Interactive Lifelog Search Engine at LSC 2020." In Proceedings of the Third Annual Workshop on Lifelog Search Challenge, pp. 57-62. 2020.
- [4] Ionescu, Bogdan, Henning Müller, Renaud Péteri, Duc-Tien Dang-Nguyen, **Liting Zhou**, Luca Piras, Michael Riegler et al. "ImageCLEF 2020: Multimedia Retrieval in Lifelogging, Medical, Nature, and Internet Applications." In European Conference on Information Retrieval, pp. 533-541. Springer, Cham, 2020.
- [5] **Zhou Liting**, Jianquan Liu, Shoji Nishimura, Joseph Antony, and Cathal

- Gurrin. "Causality Inspired Retrieval of Human-object Interactions from Video." In the 2019 International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1-6. IEEE, 2019.
- [6] Gurrin, Cathal, Hideo Joho, Frank Hopfgartner, **Liting Zhou**, Van-Tu Ninh, Tu-Khiem Le, Rami Albatal, Duc-Tien Dang-Nguyen, and Graham Healy. "Advances in Lifelog Data Organisation and Retrieval at the Lifelog-3 Task." In NII Conference on Testbeds and Community for Information Access Research, pp. 16-28. Springer, Cham, 2019.
- [7] Gurrin, Cathal, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, **Liting Zhou**, Aaron Duane, Dang Nguyen et al. "Comparing approaches to interactive lifelog search at the lifelog search challenge (LSC2018)." *ITE Transactions on Media Technology and Applications* 7, no. 2 (2019): 46-59.
- [8] Dang-Nguyen, Duc-Tien, Luca Piras, Michael Riegler, Minh-Triet Tran, **Liting Zhou**, Mathias Lux, Tu-Khiem Le, Van-Tu Ninh, and Cathal Gurrin. "Overview of ImageCLEFlifelog 2019: solve my life puzzle and lifelog moment retrieval." In *CLEF2019 Working Notes*. CEUR Workshop Proceedings, vol. 2380, pp. 09-12. 2019.
- [9] Ninh, Van-Tu, Tu-Khiem Le, **Liting Zhou**, Luca Piras, Michael Riegler, Mathias Lux, Minh-Triet Tran, Cathal Gurrin, and Duc-Tien Dang-Nguyen. "LIFER 2.0: Discovering Personal Lifelog Insights using an Interactive Lifelog Retrieval System." (2019).
- [10] Ninh, Van-Tu, Tu-Khiem Le, **Liting Zhou**, Graham Healy, Kaushik Venkataraman, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Sinead Smyth, and Cathal Gurrin. "A Baseline Interactive Retrieval Engine for Visual Lifelogs at the NTCIR-14 Lifelog-3 Task." In *NII Conference on Testbeds and Community for Information Access Research*, pp. 29-41. Springer, Cham, 2019.

- [11] Le, Tu-Khiem, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Minh-Triet Tran, **Liting Zhou**, Pablo Redondo, Sinead Smyth, and Cathal Gurrin. "LifeSeeker: Interactive Lifelog Search Engine at LSC 2019." In Proceedings of the ACM Workshop on Lifelog Search Challenge, pp. 37-40. 2019.
- [12] Dang-Nguyen, Duc-Tien, Luca Piras, Michael Riegler, **Liting Zhou**, Mathias Lux, and Cathal Gurrin. "Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval." In CLEF (Working Notes). 2018.
- [13] Dang-Nguyen, Duc-Tien, Luca Piras, Michael Riegler, **Liting Zhou**, Mathias Lux, and Cathal Gurrin. "Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval." In CLEF (Working Notes). 2018.
- [14] **Liting Zhou**, Zaher Hinbarij, Dang Nguyen, Duc Tien, and Cathal Gurrin. "LIFER: an interactive lifelog retrieval system." (2018).
- [15] Dang-Nguyen, Duc-Tien, Michael Riegler, **Liting Zhou**, and Cathal Gurrin. "Challenges and opportunities within personal life archives." In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pp. 335-343. 2018.
- [16] **Liting Zhou**, Aaron Duane, Dang Nguyen, Duc Tien, and Cathal Gurrin. "DCU at the NTCIR-13 Lifelog-2 Task." NTCIR, 2017.
- [17] **Liting Zhou**, Luca Piras, Michael Riegler, Giulia Boato, Dang Nguyen, Duc Tien, and Cathal Gurrin. "Organizer team at ImageCLEFlifelog 2017: baseline approaches for lifelog retrieval and summarization." (2017).
- [18] Dang-Nguyen, Duc-Tien, Luca Piras, Michael Riegler, Giulia Boato, **Liting Zhou**, and Cathal Gurrin. "Overview of ImageCLEFlifelog 2017: Lifelog Retrieval and Summarization." In CLEF (Working Notes). 2017.

- [19] **Liting Zhou**, Luca Piras, Michael Riegler, Giulia Boato, Dang Nguyen, Duc Tien, and Cathal Gurrin. "Organizer team at ImageCLEFlifelog 2017: baseline approaches for lifelog retrieval and summarization." (2017).
- [20] Dang- Nguyen, Duc-Tien, **Liting Zhou**, Rashmi Gupta, Michael Riegler, and Cathal Gurrin. "Building a disclosed lifelog dataset: challenges, principles and processes." In Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, pp. 1-6. 2017.
- [21] Gurrin, Cathal, Hideo Joho, Frank Hopfgartner, **Liting Zhou**, and Rami 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, pp. 705-708. 2016.
- [22] **Zhou Liting**, Cathal Gurrin, Rami Albatal, Hideo Joho, and Frank Hopfgartner. "NTCIR12-Lifelog: a test collection to support collaborative benchmarking." CEUR workshop proceedings, 2016.
- [23] Duane, Aaron, Rashmi Gupta, **Liting Zhou**, and Cathal Gurrin. "Visual Insights from Personal Lifelogs: Insight at the NTCIR-12 Lifelog LIT Task." NTCIR13. June 2016.
- [24] Zhou, Lijuan Marissa, Brian Moynagh, **Liting Zhou**, Tengqi Ye and Cathal Gurrin. MemLog, an Enhanced Lifelog Annotation and Search Tool. In: The 19th International Conference on Multimedia Modelling, 5-7 Jan 2015, Sydney, Australia.

# Bibliography

- [1] J. W. Kim, J. H. Lim, S. M. Moon, and B. Jang. Collecting health lifelog data from smartwatch users in a privacy-preserving manner. *IEEE Transactions on Consumer Electronics*, 65(3):369–378, 2019.
- [2] Cathal Gurrin, Alan F. Smeaton, Zhengwei Qiu, and Aiden Doherty. Exploring the technical challenges of large-scale lifelogging. In *Proceedings of the 4th International SenseCam & Pervasive Imaging Conference on - SenseCam '13*, pages 68–75. ACM Press, 2013.
- [3] Junho Choi, Chang Choi, Hoon Ko, and Pankoo Kim. Intelligent healthcare service using health lifelog analysis. *Journal of medical systems*, 40(8):1–10, 2016.
- [4] J. Lu, X. Zheng, M. Sheng, J. Jin, and S. Yu. Efficient human activity recognition using a single wearable sensor. *IEEE Internet of Things Journal*, pages 1–1, 2020.
- [5] EP Idoga and AI Adamu. Understanding smart wearable sensors technology: Impact on human health and fitness. *Journal of Applied Sciences and Environmental Management*, 24(7):1261–1265, 2020.
- [6] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. Lifelogging: Personal big data. *Found. Trends Inf. Retr.*, 8(1):1–125, June 2014.
- [7] F Milton, N Muhlert, C. R. Butler, A Smith, A Benattayallah, and Az Zeman. An fMRI study of long-term everyday memory using SenseCam. *Memory*, 19(7):733–744, 2011.

- [8] Philip J. Barnard, Fionnuala C. Murphy, Maria Teresa Carthery-Goulart, Cristina Ramponi, and Linda Clare. Exploring the basis and boundary conditions of SenseCam-facilitated recollection. *Memory*, 19(7):758–767, 2011.
- [9] Emma Berry, Narinder Kapur, Lyndsay Williams, Steve Hodges, Peter Watson, Gavin Smyth, James Srinivasan, Reg Smith, Barbara Wilson, and Ken Wood. The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: a preliminary report. *Neuropsychological rehabilitation*, 17(4-5):582–601, 2007.
- [10] Morgan Harvey, Marc Langheinrich, and Geoff Ward. Remembering through lifelogging: A survey of human memory augmentation. *Pervasive and Mobile Computing*, 27:14–26, 2016.
- [11] L. N. Signal, J. Stanley, M. Smith, M. B. Barr, T. J. Chambers, J. Zhou, A. Duane, C. Gurrin, A. F. Smeaton, C. McKerchar, A. L. Pearson, J. Hoek, G. L. S. Jenkin, and C. Ni Mhurchu. Children’s everyday exposure to food marketing: an objective analysis using wearable cameras. *International Journal of Behavioral Nutrition and Physical Activity*, 14(1):137, oct 2017.
- [12] Gemma Wilson, Derek Jones, Patricia Schofield, and Denis J. Martin. The use of a wearable camera to explore daily functioning of older adults living with persistent pain: Methodological reflections and recommendations. *Journal of Rehabilitation and Assistive Technologies Engineering*, 5:205566831876541, 2018.
- [13] Thi Hoa Cuc Nguyen, Jean Christophe Nebel, and Francisco Florez-Revuelta. Recognition of activities of daily living with egocentric vision: A review. *Sensors*, 16(1), 2016.
- [14] Bethan Everson, Kelly A. Mackintosh, Melitta A. McNarry, Charlotte Todd, and Gareth Stratton. Can Wearable Cameras be Used to Validate School-Aged Children’s Lifestyle Behaviours? *Children*, 6(2):20, 2019.

- [15] Qianling Zhou, Di Wang, Cliona Ni Mhurchu, Cathal Gurrin, Jiang Zhou, Yu Cheng, and Haijun Wang. The use of wearable cameras in assessing children’s dietary intake and behaviours in China. *Appetite*, 2019.
- [16] Roberto Hoyle, Robert Templeman, Steven Armes, Denise Anthony, David Crandall, and Apu Kapadia. Privacy behaviors of lifeloggers using wearable cameras. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct*, pages 571–582, 2014.
- [17] J. Gemmell, G. Bell, and R. Lueder. MyLifeBits: a personal database for everything. *Communications of the ACM*, 49(1):88–95, 2006.
- [18] Gordon Bell and Jim Gemmell. A digital life. *Scientific American*, 2007.
- [19] Steve Mann, Kris M Kitani, Yong Jae Lee, MS Ryoo, and Alireza Fathi. An introduction to the 3rd workshop on egocentric (first-person) vision. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 827–832. IEEE, 2014.
- [20] Vannevar Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, Jul 1945.
- [21] Theodor H. Nelson. Managing immense storage. *BYTE*, 13(1):225–238, January 1988.
- [22] C Bell and Jim Gemmell. Total recall: how the e-memory revolution will change everything: Dutton. *Berry, E., Kapur, N., Williams, L., Hodges, S., Watson, P., Smyth, G., et al.(2007). The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: a preliminary report. Neuropsychological Rehabilitation*, 17(45):582601, 2009.

- [23] S. Mann. Wearable computing: A first step toward personal imaging. *Computer*, 30(2):25–32, 1997.
- [24] David H. Nguyen, Gabriela Marcu, Gillian R. Hayes, Khai N. Truong, James Scott, Marc Langheinrich, and Christof Roduner. Encountering sensecam: personal recording technologies in everyday life. In *UbiComp '09: Proceedings of the 11th international conference on Ubiquitous computing*, pages 165–174, 2009.
- [25] Duc-Tien Dang-Nguyen, Liting Zhou, Rashmi Gupta, Michael Riegler, and Cathal Gurrin. Building a disclosed lifelog dataset: challenges, principles and processes. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, page 22. ACM, 2017.
- [26] Abigail J. Sellen and Steve Whittaker. Beyond total capture: a constructive critique of lifelogging. *Comm. ACM*, 53(5):70–77, 2010.
- [27] Aiden R. Doherty, Chris J.A. Moulin, and Alan F. Smeaton. Automatically assisting human memory: A sensecam browser. *Memory*, 7(19):785–795, 2011.
- [28] Aiden Roger Doherty, Niamh Caprani, Ciaran O Conaire, Vaiva Kalnikaite, Cathal Gurrin, Alan F. Smeaton, and Noel E. O Connor. Passively recognising human activities through lifelogging. *Computers in Human Behavior*, 27(5):1948–1958, 2011.
- [29] Ramesh Jain and Pinaki Sinha. Content without context is meaningless. In *Proceedings of the international conference on Multimedia*, MM '10, pages 1259–1268, New York, NY, USA, 2010. ACM.
- [30] A R Doherty, K Pauly-Takacs, N Caprani, C Gurrin, C J A Moulin, N E O'Connor, and A F Smeaton. Experiences of aiding autobiographical memory using the sensecam. *Human-Computer Interaction*, 27(1-2):151–174, 2012.



- [31] Rashmi Gupta and Cathal Gurrin. Considering manual annotations in dynamic segmentation of multimodal lifelog data. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 34–39. IEEE, 2019.
- [32] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, and Rami Albatal. NTCIR lifelog: The first test collection for lifelog research. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16*, pages 705–708, 2016.
- [33] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Rashmi Gupta, Rami Albatal, and Duc-Tien Dang-Nguyen. Overview of NTCIR-13 lifelog-2 task. In *Proceedings of NTCIR-13, Tokyo, Japan*, pages 6–11, 2017.
- [34] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Van-Tu Ninh, Tu-Khiem Le, Rami Albatal, Duc-Tien Dang-Nguyen, and Graham Healy. Overview of NTCIR-14 lifelog-3 task. In *Proceedings of NTCIR-14, Tokyo, Japan*, 2019.
- [35] Martin Dodge and Rob Kitchin. “Outlines of a world coming into existence”: Pervasive computing and the ethics of forgetting. *Environment and Planning B*, 34(3):431–445, 2007.
- [36] Minh-Son Dao, Duc-Tien Dang-Nguyen, Michael Riegler, and Cathal Gurrin. Smart Lifelogging: Recognizing Human Activities using PHASOR. In *International Conference on Pattern Recognition Applications and Methods*, pages 761–767, 2017.
- [37] Melanie Swan. The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data*, 1(2):85–89, June 2013.
- [38] Kieron O’Hara, Mischa M. Tuffield, and Nigel Shadbolt. Lifelogging: Privacy and empowerment with memories for life. *Identity in the Information Society*, 1(1):155–172, 2008.

- [39] Jim Gemmell, Gordon Bell, Roger Lueder, Steven Drucker, and Curtis Wong. Mylifebits: Fulfilling the memex vision. In *Proceedings of the Tenth ACM International Conference on Multimedia*, ACM Multimedia '02, pages 235–238, New York, NY, USA, 2002. ACM.
- [40] C. Gurrin, Hideo Joho, F. Hopfgartner, L. Zhou, and Rami Albatal. Overview of ntcir-12 lifelog task. In *NTCIR*, 2016.
- [41] Cathal Gurrin, Rami Albatal, Hideo Joho, and Kaori Ishii. A privacy by design approach to lifelogging. 2014.
- [42] Steve Hodges, Emma Berry, and Ken Wood. SenseCam: A wearable camera that stimulates and rehabilitates autobiographical memory. *Memory*, 7(19):685–696, 2011.
- [43] A.R. Doherty and A.F. Smeaton. Automatically segmenting lifelog data into events. In *Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, pages 20–23. IEEE, 2008.
- [44] Daniel Kahneman, Alan B. Krueger, David A. Schkade, Norbert Schwarz, and Arthur A. Stone. A survey method for characterizing daily life experience: The day reconstruction method. *Science*, 306:1776–1780, 2004.
- [45] Ling Bao and Stephen Intille. Activity recognition from user-annotated acceleration data. In Alois Ferscha and Friedemann Mattern, editors, *Pervasive Computing*, volume 3001 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin / Heidelberg, 2004.
- [46] Daragh Byrne, Barry Lavelle, Aiden R. Doherty, Gareth J.F. Jones, and Alan F. Smeaton. Using bluetooth and gps metadata to measure event similarity in sensecam images. *5th International Conference on Intelligent Multimedia and Ambient Intelligence*, July 2007.

- [47] Aiden R Doherty and Alan F Smeaton. Automatically augmenting lifelog events using pervasively generated content from millions of people. *Sensors (Peterborough)*, 10(3):1423–1446, 2010.
- [48] Henk M Blanken, Arjen P de Vries, Henk Ernst Blok, and Ling Feng. *Multimedia retrieval*. Springer, 2007.
- [49] Charles B. Callaway and James C. Lester. Narrative prose generation. *Artif. Intell.*, 139(2):213–252, August 2002.
- [50] Daragh Byrne, Aisling Kelliher, and Gareth JF Jones. Life editing: third-party perspectives on lifelog content. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1501–1510, 2011.
- [51] James Allan. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media, 2012.
- [52] Le Zhao, Min Zhang, and Shaoping Ma. The nature of novelty detection. *Inf. Retr.*, 9(5):521–541, November 2006.
- [53] Y. Ogras and Hakan Ferhatosmanoglu. Online summarization of dynamic time series data. *The VLDB Journal*, 15(1):84–98, January 2006.
- [54] D. Tancharoen, T. Yamasaki, and K. Aizawa. Practical life log video indexing based on content and context. In *Multimedia Content Analysis, Management, and Retrieval*, 2006.
- [55] Zhengwei Qiu, A.R. Doherty, C. Gurrin, and A.F. Smeaton. Mining user activity as a context source for search and retrieval. In *Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on*, pages 162 –166, june 2011.
- [56] S. Lee, G. Gong, I. Hwang, and S. g. Lee. Lifelogon: A practical lifelog system for building and exploiting lifelog ontology. In *2010 IEEE International*

*Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, pages 367–373, June 2010.

- [57] *LSC '18: Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, New York, NY, USA, 2018. Association for Computing Machinery.
- [58] Duc Tien Dang Nguyen, Luca Piras, Michael Riegler, Giulia Boato, Liting Zhou, and Cathal Gurrin. Overview of imageclef lifelog 2017: lifelog retrieval and summarization. 2017.
- [59] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. Overview of ImageCLEFlifelog 2018: daily living understanding and lifelog moment retrieval. In *CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France (September 10-14 2018)*, 2018.
- [60] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Minh-Triet Tran, Liting Zhou, Mathias Lux, Tu-Khiem Le, Van-Tu Ninh, and Cathal Gurrin. Overview of ImageCLEFlifelog 2019: Solve my life puzzle and Lifelog Moment Retrieval. In *CLEF2019 Working Notes, CEUR Workshop Proceedings, Lugano, Switzerland, September 09-12 2019*. CEUR-WS.org <<http://ceur-ws.org>>.
- [61] Lei Lei, Xuguang Ren, Nigel Franciscus, Junhu Wang, and Bela Stantic. Event prediction based on causality reasoning. In *Asian Conference on Intelligent Information and Database Systems*, pages 165–176. Springer, 2019.
- [62] Humayun Kayesh, Md Islam, Junhu Wang, et al. On event causality detection in tweets. *arXiv preprint arXiv:1901.03526*, 2019.
- [63] Manish Sharma and Rahul Patel. A survey on information retrieval models, techniques and applications. *International Journal of Emerging Technology and Advanced Engineering*, 3(11):542–545, 2013.

- [64] Jitendra Nath Singh and Sanjay Kumar Dwivedi. Analysis of vector space model in information retrieval. In *National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC*, 2012.
- [65] D. Hiemstra and De A. Vries. Relating the new language models of information retrieval to the traditional retrieval models. *CTIT technical report series*, pages 1–14, 2000.
- [66] Hyowon Lee, Alan F. Smeaton, Noel E. O’Connor, Gareth Jones, Michael Blighe, Daragh Byrne, Aiden Doherty, and Cathal Gurrin. Constructing a SenseCam visual diary as a media process. *Multimedia Systems*, 14(6):341–349, 2008.
- [67] Gabriel de Oliveira Barra, Alejandro Cartas Ayala, Marc Bolaños, Mariella Dimiccoli, Maedeh Aghaei, Marc Carné, Xavier Giro-I-Nieto, and Petia Radeva. LEMoRe: A lifelog engine for moments retrieval at the NTCIR-lifelog LSAT task. In *Proceedings of NTCIR-12, Tokyo, Japan*, 2016.
- [68] S. Chowdhury, M. S. Ferdous, and J. M. Jose. A user-study examining visualization of lifelogs. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, June 2016.
- [69] Aiden R. Doherty, Katalin Pauly-Takacs, Niamh Caprani, Cathal Gurrin, Chris J. a. Moulin, Noel E. O’Connor, and Alan F. Smeaton. Experiences of Aiding Autobiographical Memory Using the SenseCam. *Human-Computer Interaction*, 27(1-2):151–174, 2012.
- [70] Ellen M. Voorhees and Angela Ellis, editors. *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*, volume Special Publication 500-324. National Institute of Standards and Technology (NIST), 2017.
- [71] Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier, editors. *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*,

*Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

- [72] Yiqun Liu, Makoto P. Kato, Charles L. A. Clarke, Noriko Kando, and Tetsuya Sakai. Report on NTCIR-13: the thirteenth round of NII testbeds and community for information access research. *SIGIR Forum*, 52(1):102–110, 2018.
- [73] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew Lungren, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. Overview of ImageCLEF 2018: Challenges, Datasets and Evaluation. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 309–334, Cham, 2018. Springer International Publishing.
- [74] Martha Larson, Piyush Arora, Claire-Hélène Demarty, Michael Riegler, Benjamin Bischke, Emmanuel Dellandréa, Mathias Lux, Alastair Porter, and Gareth J. F. Jones, editors. *Working Notes Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29-31 October 2018*, volume 2283 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.
- [75] Duc-Tien Dang-Nguyen, Michael Riegler, Liting Zhou, and Cathal Gurrin. Challenges and opportunities within personal life archives. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 335–343. ACM, 2018.
- [76] Aaron Duane, Liting Zhou, Duc-Tien Dang-Nguyen, and Cathal Gurrin. DCU at the NTCIR-13 lifelog-2 task. In *Proceedings of NTCIR-13, Tokyo, Japan*,

2017.

- [77] Min-Huan Fu, Chang Chia-Chun, Gen-Hsen Huang, and Hsin-His Chen. Introducing external textual knowledge for lifelog retrieval and annotation. In *Proceedings of NTCIR-14, Tokyo, Japan*, 2019.
- [78] Van-Tu Ninh, Tu-Khiem Le, Liting Zhou, Graham Healy, Kaushik Venkataraman, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Sinead Smith, and Cathal Gurrin. A baseline interactive retrieval engine for the NTICR-14 Lifelog-3 semantic access task. In *Proceedings of NTCIR-14, Tokyo, Japan*, 2019.
- [79] Isadora Van Khan Nguyen, Pranita Shrestha, Min Zhang, Yiqun Liu, and Shaoping Ma. THUIR at the NTCIR-14 lifelog-3 task: How does lifelog help the user’s status recognition. In *Proceedings of NTCIR-14, Tokyo, Japan*, 2019.
- [80] Manfred Jürgen Primus, Bernd Münzer, Andreas Leibetseder, and Klaus Schoeffmann. The itec collaborative video search system at the video browser showdown 2018. In Klaus Schoeffmann, Thanarat H. Chalidabhongse, Chong Wah Ngo, Supavadee Aramvith, Noel E. O’Connor, Yo-Sung Ho, Moncef Gabbouj, and Ahmed Elgammal, editors, *MultiMedia Modeling*, pages 438–443, Cham, 2018. Springer International Publishing.
- [81] Bernd Münzer, Andreas Leibetseder, Sabrina Kletz, Manfred Jürgen Primus, and Klaus Schoeffmann. lifexplore at the lifelog search challenge 2018. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge, LSC ’18*, pages 3–8, New York, NY, USA, 2018. ACM.
- [82] Adrià Alsina, Xavier Giró, and Cathal Gurrin. An interactive lifelog search engine for lsc2018. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge, LSC ’18*, pages 30–32, New York, NY, USA, 2018. ACM.
- [83] Aaron Duane, Cathal Gurrin, and Wolfgang Hürst. Virtual reality lifelog explorer: Lifelog search challenge at acm icmr 2018. In *Proceedings of the*

- 2018 ACM Workshop on The Lifelog Search Challenge*, LSC '18, pages 20–23, New York, NY, USA, 2018. ACM.
- [84] Thanh-Dat Truong, Tung Dinh-Duy, Vinh-Tiep Nguyen, and Minh-Triet Tran. Lifelogging retrieval based on semantic concepts fusion. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, LSC '18, pages 24–29, New York, NY, USA, 2018. ACM.
  - [85] Viet-Khoa Vo-Ho, Quoc-An Luong, Duy-Tam Nguyen, Mai-Khiem Tran, and Minh-Triet Tran. Personal diary generation from wearable cameras with concept augmented image captioning and wide trail strategy. In *Proceedings of the Ninth International Symposium on Information and Communication Technology*, SoICT 2018, pages 367–374, New York, NY, USA, 2018. ACM.
  - [86] Jakub Lokoč, Tomáš Souček, and Gregor Kovalčík. Using an interactive video retrieval tool for lifelog data. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, LSC '18, pages 15–19, New York, NY, USA, 2018. ACM.
  - [87] Ana Garcia del Molino, Bappaditya Mandal, Jie Lin, Joo Hwee Lim, Vigneshwaran Subbaraju, and Vijay Chandrasekhar. Vc-i2r@ imageclef2017: Ensemble of deep learned features for lifelog video summarization. 2017.
  - [88] Mihai Dogariu and Bogdan Ionescu. A textual filtering of hog-based hierarchical clustering of lifelog data. In *CLEF (Working Notes)*, 2017.
  - [89] Mihai Dogariu and Bogdan Ionescu. Multimedia Lab @ CAMPUS at ImageCLEFlifelog 2018 Lifelog Moment Retrieval, September 10-14 2018.
  - [90] Ergina Kavallieratou, Carlos R. del Blanco, Carlos Cuevas, and Narciso García. Retrieving Events in Life Logging, September 10-14 2018.
  - [91] Tsun-Hsien Tang, Min-Huan Fu1, Hen-Hsen Huang, Kuan-Ta Chen, and Hsin-Hsi Chen. NTU NLP-Lab at ImageCLEFlifelog 2018: Visual Concept Selection



with Textual Knowledge for Understanding Activities of Daily Living and Life Moment Retrieval, September 10-14 2018.

- [92] Minh-Triet Tran, Thanh-Dat Truong, Tung Dinh-Duy, Viet-Khoa Vo-Ho, Quoc-An Luong, and Vinh-Tiep Nguyen. Lifelog Moment Retrieval with Visual Concept Fusion and Text-based Query Expansion, September 10-14 2018.
- [93] Fatma Ben Abdallah, Ghada Feki, Mohamed Ezzarka, Anis Ben Ammar, and Chokri Ben Amar. Regim Lab Team at ImageCLEFlifelog LMRT Task 2018, September 10-14 2018.
- [94] Fatma Ben Abdallah, Ghada Feki, Anis Ben Ammar, and Chokri Ben Amar. Big data for lifelog moments retrieval improvement. In *CLEF (Working Notes)*, 2019.
- [95] Mihai Dogariu and Bogdan Ionescu. Multimedia lab@ imageclef 2019 lifelog moment retrieval task. In *CLEF (Working Notes)*, 2019.
- [96] R Ribeiro, AJR Neves, and JL Oliveira. Uapbioinformatics working notes at imageclef 2019 lifelog moment retrieval (lmrt) task. In *CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org* < <http://ceur-ws.org>, 2019.
- [97] Stefan Taubert, Stefan Kahl, Danny Kowerko, and Maximilian Eibl. Automated lifelog moment retrieval based on image segmentation and similarity scores. In *CLEF (Working Notes)*, 2019.
- [98] Nguyen-Khang Le, Dieu-Hien Nguyen, Vinh-Tiep Nguyen, and Minh-Triet Tran. Lifelog moment retrieval with advanced semantic extraction and flexible moment visualization for exploration. In *CLEF (Working Notes)*, 2019.
- [99] Minh-Son Dao, Anh-Khoa Vo, Trong-Dat Phan, and Koji Zettsu. Bidal@ imagecleflifelog2019: The role of content and context of daily activities in insights from lifelogs. In *CLEF (Working Notes)*, 2019.

- [100] Pengfei Zhou, Cong Bai, and Jie Xia. Zjutcvr team at imagecleflifelog2019 lifelog moment retrieval task. In *CLEF (Working Notes)*, 2019.
- [101] Maxime Tournadre, Guillaume Dupont, Vincent Pauwels, Bezeid Cheikh Mohamed Lmami, and Alexandru-Lucian Gînsca. A multimedia modular approach to lifelog moment retrieval. In *CLEF (Working Notes)*, 2019.
- [102] Li Deng and Dong Yu. *Deep Learning*. Now Publishers Incorporated, 2014.
- [103] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.
- [104] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [105] G. E. Hinton. Deep belief networks. *Scholarpedia*, 4(5):5947, 2009. revision #91189.
- [106] B. Hutchinson, L. Deng, and D. Yu. Tensor deep stacking networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1944–1957, 2013.
- [107] Keiron O’Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.
- [108] Shulin Yang. *Feature Engineering in Fine-Grained Image Classification*. PhD Thesis, University of Washington, 2013.
- [109] Honglak Lee. *Unsupervised feature learning via sparse hierarchical representations*. PhD thesis, Stanford University, 2010.
- [110] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the*

- IEEE conference on computer vision and pattern recognition*, pages 4594–4602, 2016.
- [111] A. C. H. Choong and N. K. Lee. Evaluation of convolutionary neural networks modeling of dna sequences using ordinal versus one-hot encoding method. In *2017 International Conference on Computer and Drone Applications (IConDA)*, pages 60–65, 2017.
  - [112] Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. Cross-lingual information retrieval with bert. *arXiv preprint arXiv:2004.13005*, 2020.
  - [113] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. Bert representations for video question answering. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1556–1565, 2020.
  - [114] Ethan A Chi, John Hewitt, and Christopher D Manning. Finding universal grammatical relations in multilingual bert. *arXiv preprint arXiv:2005.04511*, 2020.
  - [115] Alba Herrera-Palacio, C. Ventura, Carina Silberer, Ionut-Teodor Sorodoc, Gemma Boleda, and Xavier Giro i Nieto. Recurrent instance segmentation using sequences of referring expressions. *ArXiv*, abs/1911.02103, 2019.
  - [116] Fei-Fei Li, Andrej Karpathy, and Justin Johnson. *CS231n: Convolutional Neural Networks for Visual Recognition*, 2016. Available at <http://cs231n.github.io/convolutional-networks/>.
  - [117] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *International conference on medical image computing and computer-assisted intervention*, pages 246–253. Springer, 2013.

- [118] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011.
- [119] Rebecca Mason and Eugene Charniak. Nonparametric method for data-driven image captioning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, 2014.
- [120] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233, 2014.
- [121] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.
- [122] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.
- [123] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.
- [124] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010.
- [125] Polina Kuznetsova, Vicente Ordonez, Tamara L Berg, and Yejin Choi. Treetalk: Composition and compression of trees for image descriptions.

- Transactions of the Association for Computational Linguistics*, 2:351–362, 2014.
- [126] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.
  - [127] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *CoRR*, abs/1505.04474, 2015.
  - [128] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389. IEEE, 2018.
  - [129] Bingjie Xu, Junnan Li, Yongkang Wong, Mohan S. Kankanhalli, and Qi Zhao. Interact as you intend: Intention-driven human-object interaction detection, 2018.
  - [130] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018.
  - [131] J.W. Creswell. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Pearson, 2012.
  - [132] Cathal Gurrin, Xavier Giro-i Nieto, Petia Radeva, Mariella Dimiccoli, Håvard Johansen, Hideo Joho, and Vivek K Singh. Lta 2016: The first workshop on lifelogging tools and applications. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1487–1488, 2016.
  - [133] Marc Langheinrich. Privacy by design—principles of privacy-aware ubiquitous systems. In *International conference on Ubiquitous Computing*, pages 273–291. Springer, 2001.

- [134] Zaher Hinbarji, Rami Albatal, Noel O'Connor, and Cathal Gurrin. Loggerman, a comprehensive logging and visualization tool to capture computer usage. In *International Conference on Multimedia Modeling*, pages 342–347. Springer, 2016.
- [135] Cathal Gurrin, Daragh Byrne, Noel O'Connor, Gareth JF Jones, and Alan F. Smeaton. Architecture and challenges of maintaining a large-scale, context-aware human digital memory. In *International Conference on Visual Information Engineering*, pages 158–163, 2008.
- [136] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Helbert Arenas, Giulia Boato, Duc-Tien Dang-Nguyen, Yashin Dicente Cid, Carsten Eickhoff, Alba G Seco de Herrera, Cathal Gurrin, et al. Overview of imageclef 2017: Information extraction from images. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 315–337. Springer, 2017.
- [137] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba García Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew Lungren, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), Avignon, France, September 10-14 2018. LNCS Lecture Notes in Computer Science, Springer.
- [138] Bogdan Ionescu, Henning Müller, Renaud Péteri, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Dzmitri Klimuk, Aleh Tarasau, Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Minh-Triet Tran, Mathias Lux,

- Cathal Gurrin, Obioma Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, Narciso Garcia, Ergina Kavallieratou, Carlos Roberto del Blanco, Carlos Cuevas Rodríguez, Nikos Vasillopoulos, Konstantinos Karampidis, Jon Chamberlain, Adrian Clark, and Antonio Campello. ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), Lugano, Switzerland, September 9-12 2019. LNCS Lecture Notes in Computer Science, Springer.
- [139] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Liting Zhou, Aaron Duane, Andreas Leibetseder, Michael Riegler, and Luca Piras. Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge ( LSC2018 ). *ITE Transactions on Media Technology and Applications*, 7(2):46–59, 2019.
- [140] Cathal Gurrin, Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Björn Þór Jónsson, Jakub Lokoš, Wolfgang Hürst, Minh-Triet Tran, and Klaus Schoeffmann. Introduction to the third annual lifelog search challenge (lsc’20). In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 584–585, 2020.
- [141] Liting Zhou, Luca Piras, Michael Riegler, Giulia Boato, Duc Tien Dang Nguyen, and Cathal Gurrin. Organizer team at imagecleflifelog 2017: baseline approaches for lifelog retrieval and summarization. *CLEF working notes, CEUR*, 2017.
- [142] Liting Zhou, Duc-Tien Dang-Nguyen, and Cathal Gurrin. A baseline search engine for personal life archives. In *Proceedings of the 2nd Workshop on Lifelogging Tools and Applications*, LTA ’17, pages 21–24, New York, NY, USA, 2017. ACM.

- [143] Duc-Tien Dang-Nguyen, Luca Piras, Giorgio Giacinto, Giulia Boato, and Francesco GB De Natale. Multimodal retrieval with diversification and relevance feedback for tourist attraction images. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2017. Accepted.
- [144] J.-T. Huang, C.-H. Shen, S.-M. Phoong, and H. Chen. Robust measure of image focus in the wavelet domain. In *Intelligent Signal Processing and Communication Systems*, pages 157–160, 2005.
- [145] Duc-Tien Dang-Nguyen, Luca Piras, Giorgio Giacinto, Giulia Boato, and Francesco GB De Natale. A hybrid approach for retrieving diverse social images of landmarks. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2015.
- [146] Liting Zhou, Zaher Hinbarji, Duc-Tien Dang-Nguyen, and Cathal Gurrin. Lifer: An interactive lifelog retrieval system. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, pages 9–14. ACM, 2018.
- [147] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.
- [148] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS’15, pages 91–99, Cambridge, MA, USA, 2015. MIT Press.
- [149] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.



- [150] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [151] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.
- [152] Amy Fire and Song-Chun Zhu. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.
- [153] Chao-Han Huck Yang, Yi-Chieh Liu, Pin-Yu Chen, Xiaoli Ma, and Yi-Chang James Tsai. When causal intervention meets image masking and adversarial perturbation for deep neural networks. *arXiv preprint arXiv:1902.03380*, 2019.
- [154] Yunchuan Sun, Yufeng Shi, and Zhengjun Zhang. Finance big data: management, analysis, and applications, 2019.
- [155] Barbara Bedowska-Sojka and Agata Kliber. The causality between liquidity and volatility in the polish stock market. *Finance Research Letters*, 30:110–115, 2019.
- [156] Jundong Li, Osmar R Zaiane, and Alvaro Osornio-Vargas. Discovering statistically significant co-location rules in datasets with extended spatial objects. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 124–135. Springer, 2014.
- [157] Marie-Abèle Bind. Causal modeling in environmental health. *Annual review of public health*, 40:23–43, 2019.
- [158] Made Ary Meitriana, I Kadek Rai Suwena, and Luh Indrayani. The influence of fraud triangle and theory of planned behavior on students academic fraud

- in bali. In *International Conference on Tourism, Economics, Accounting, Management, and Social Science (TEAMS 2018)*. Atlantis Press, 2019.
- [159] Michael D Feters and José F Molina-Azorin. In this issue: Innovations in mixed methods—causality, case study research with a circular joint display, social media, grounded theory, and phenomenology, 2019.
- [160] Karthir Prabhakar, Sangmin Oh, Ping Wang, Gregory D Abowd, and James M Rehg. Temporal causality for the analysis of visual events. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1967–1974. IEEE, 2010.
- [161] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016.
- [162] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. Cascaded human-object interaction recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4263–4272, 2020.
- [163] Kaisheng Xu, Hanli Wang, and Pengjie Tang. Image captioning with deep lstm based on sequential residual. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*, pages 361–366. IEEE, 2017.
- [164] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.
- [165] Anna Zaremba and Tomaso Aste. Measures of causality in complex datasets with application to financial data. *Entropy*, 16(4):2309–2349, 2014.

- [166] Steven L Bressler and Anil K Seth. Wiener-granger causality: a well established methodology. *Neuroimage*, 58(2):323–329, 2011.
- [167] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1017–1025, 2015.
- [168] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.
- [169] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [170] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008, 2017.
- [171] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [172] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.
- [173] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.
- [174] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

- [175] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [176] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.
- [177] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.
- [178] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.
- [179] Liting Zhou, Zaher Hinbarji, Duc-Tien Dang-Nguyen, and Cathal Gurrin. LIFER: An Interactive Lifelog Retrieval System. In *Workshop on Lifelog Search Challenge*, 2018. Accepted.
- [180] Cathal Gurrin, Xavier Giro-i Nieto, Pedia Radeva, Mariella Dimiccoli, Duc-Tien Dang-Nguyen, and Hideo Joho. LTA 17: The second Workshop on Lifelogging Tools and Applications. In *Proceedings of the 2nd Workshop on Lifelogging Tools and Applications*. ACM, 2017.
- [181] J. Xu, S. Kemeny, G. Park, C. Frattali, and A. Braun. Language in context: emergent features of word, sentence, and narrative comprehension. *Neuroimage*, 25(3):1002–1015, 2005.
- [182] J. Gemmell, A. Aris, and R. Lueder. Telling stories with MyLifeBits. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1536–1539, 2005.

- [183] Liting Zhou, Zaher Hinbarji, and Cathal Gurrin. DCU Team at LSC : Interactive Retrieval from Lifelog Archives. *ACM ICMR 2018*, pages 15–18, 2018.
- [184] VT Ninh, TK Le, L Zhou, L Piras, M Riegler, M Lux, MT Tran, C Gurrin, and DT Dang-Nguyen. Lifer 2.0: discover personal lifelog insight by interactive lifelog retrieval system. In *CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org* < <http://ceur-ws.org>, 2019.
- [185] Quoc V. Le. *Scalable feature learning*. PhD Thesis, Stanford University, 2013.
- [186] Mehmet Gönen and Ethem Alpaydm. Multiple kernel learning algorithms. *Journal of machine learning research*, 12(Jul):2211–2268, 2011.
- [187] Shulin Yang. *Feature Engineering in Fine-Grained Image Classification*. PhD Thesis, University of Washington, 2013.
- [188] Hao Wang, Doyen Sahoo, Chenghao Liu, Ke Shu, Palakorn Achananuparp, Ee-peng Lim, and Steven CH Hoi. Cross-modal food retrieval: Learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *arXiv preprint arXiv:2003.03955*, 2020.
- [189] S. Chung, J. S. Chung, and H. Kang. Perfect match: Self-supervised embeddings for cross-modal retrieval. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):568–576, 2020.