Voxento: A Prototype Voice-controlled Interactive Search Engine for Lifelogs

Ahmed Alateeq
Ahmed.Alateeq2@mail.dcu.ie
School of Computing
Dublin City University
Ireland

Mark Roantree
Mark.Roantree@dcu.ie
Insight Centre for Data Analytics
Dublin City University
Ireland

Cathal Gurrin
Cathal.Gurrin@dcu.ie
School of Computing
Dublin City University
Ireland

ABSTRACT

In this paper, we describe an interactive voice-based retrieval system for lifelogs which has been developed to participate in the third Lifelog Search Challenge LSC'20, at ACM ICMR'20. Based on a standard text-based retrieval methodology, the novelty of Voxento is in the interactive voice facility that allows a user to interact with a personal lifelog using simple voice commands. Voxento was developed as an initial prototype of a pervasive computing system that can be deployed for wearable technologies such as Google Glass. The version of Voxento described in this paper has been optimised for use with a desktop computer in order to be competitive at the LSC'20 challenge.

CCS CONCEPTS

• **Information systems → Search interfaces**; **Multimedia DBs**;
• **Human-centered computing → Sound-based input/output**.

KEYWORDS

Interactive retrieval; lifelog; voice interaction; Speech Recognition

ACM Reference Format:
Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2020. Voxento: A Prototype Voice-controlled Interactive Search Engine for Lifelogs. In *Proceedings of the Third Annual Workshop on the Lifelog Search Challenge (LSC '20), June 9, 2020, Dublin, Ireland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3379172.3391728

1 INTRODUCTION

Lifelogging is the action of digitising one's everyday life activities with the aid of technology including electronic devices and computer applications. The lifelogging phenomenon has become more popular in recent years due to ever-decreasing cost of computing devices, data storage [14] and the availability of software for processing and analysing digital data [20]. This has lead to some forms of lifelogging becoming a normative activity, such as those represented by the quantified-self movement [18], and there have been a number of studies highlighting the potential health benefits of lifelogging [6, 23]. In our research, we share a common definition with [8] which describes lifelogging as a 'form of pervasive computing, consisting of a unified digital record of the totality of an individual's experiences' which is gathered as multimodal data from various wearable, fixed and software sensors.

As a result of the large volume of data that can be generated from lifelogging, it becomes possible to gain advance into an individual's life experience [14], as has been achieved in large-scale epidemiological studies [25] using wearable camera data to understand populations. In order to be effective as assistive technology or as a tool for researching populations, there needs to be a suite of analytics and retrieval tools that facilitate rigorous interrogation of lifelog archives. Hence, the Lifelog Search Challenge is a participation workshop which invites researchers to develop interactive lifelog retrieval engines and compare then in a live competition incorporating both expert and novice users [13].

In this paper, we describe a prototype voice-controlled interactive retrieval system for lifelogs. While there has been a significant research effort into off-line and interactive lifelog retrieval, there had been no voice-controlled interactive retrieval engines, to the best of our knowledge. Hence, this challenge motivates the research into the feasibility of controlling a retrieval system by voice. We introduce Voxento, which uses the Google Web Speech API to detect a spoken query and recognise specific commands such as submit a search task or stop recording. In turn, Voxento provides both a visual and a voice-based response for users. The system is easy to use and has clear set of interface commands. In the following sections, we present the architecture of Voxento; describe its components; the datasets used in our validation; and describe the interactive interface developed for the Lifelog Search Challenge 2020. As this system faced new challenges, we feel it opens the door to further research and experiments as one seeks to improve the system's capabilities.

2 RELATED WORK

Since the development of the first interactive lifelog retrieval system MyLifeBits, researchers have developed several retrieval systems involving various features and tools used to enhance exploration, interaction and the efficiency of lifelog search engines [13]. As a result of the two previous LSC workshops, there were a number of retrieval systems which were designed to enhance information retrieval and support multiple inputs or search tasks such as query-by-example, [19] and [22]. However, no system uses speech recognition or speech synthesis in the context of lifelogging. Notably, Voxento not only converts speech to text but it also supports commands to do further actions by voice in an interactive way.

LSC '20, June 9, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7136-0/20/06…$15.00
https://doi.org/10.1145/3379172.3391728

Lifelog data attracts new research and ideas on how to visualise and explore data in an interactive way to enhance user experience [7]. A virtual reality retrieval system [9] was developed and performed best in the first LSC challenge. In this work, the authors used virtual reality as a way to visualise lifelog data and perform search tasks based on a temporal filter and concepts selection. In other research [15], the authors presented a new feature with virtual reality where users can see photographs located on a world map based on geographic data using a head-mounted display. However, large volumes of data explored in a limited screen size make visualisation difficult and detecting a relevant item could take a long time. Researchers in [21] evaluated different platforms for accessing lifelog data and discovered that the limitations of display technologies affect lifelog retrieval performance. The Voxento system is a flexible and web-based application that can be accessed from any computer with a connection to the internet and audio capabilities.

In terms of data access or query input, which is the focus of our system, most queries are long and complex with targeted search [7] and consists of numbers (dates) and strings (places or objects). An example would be like *"find a moment when I was eating dinner with my family in a pizza restaurant, the time was about 10:15 and I think it was Thursday or Friday"*. We argue that speaking the query will be faster and easier than regular textual input. Also, speaking, bring a more natural communication medium, could have the benefit of facilitating richer query descriptions than would be expected when typing queries.

The concept of recording voice as a means of detecting speech happened in 1791 when Wolfgang von Kempelen built a system called "Acoustic-Mechanical Speech Machine" [17]. The first modern system was developed to detect numbers at Bell Labs [10] for telco requirements. Since then and due to the introduction of advanced methods, models and algorithms, speech recognition has been greatly improved and become more accurate. Many recent systems were developed as a form of assistive technology, such as systems to assist patients with physical limitations to control appliances in a hospital room [2]. Another example is using speech recognition to facilitate blind individuals to manipulate user interfaces that would otherwise be inaccessible [16]. The Voxento system outlined in this paper is designed in the spirit of these assertive technologies and is instantiated as an intuitive voice interface to a lifelog. The main idea behind this work is, to be useful in everyday life, that a lifelog must provide a ubiquitous interface and not require the individual to stop what they are doing and pull-up a desktop interface to access their lifelog. It is our conjecture that such a retrieval system provides a more natural user interface to a lifelog. In this prototype implementation, the voice interface operates in conjunction with a desktop interface. In later work, we will migrate to a fully wearable platform using technologies such as Google glass.

3 SPEECH RECOGNITION

A speech recognition API facilitates users by providing direct input through voice and also, the ability to use specific and customized commands. In fact, there are a wide array of various applications which use speech recognition as an assistant such as Google Assistant or Siri in mobile platforms, car driving control, smart home devices, computer applications and web-based applications.

The first Google speech recognition was developed in 2008 and was known as Google voice search [5]. In 2013, Google then released the Web Speech API included with chrome browser version 25 and later versions. In the same year, research was published describing and measuring the performance of Google Web Speech in 2013 and found that 74% of the speech was recognized without any error although the speech recognizer had difficulties with sentences [1]. In 2017, according to a presentation in annual Internet Trends, Google speech recognition accuracy reached 95%, quite similar to the threshold of human accuracy as stated by April in Vox Media [11]. Since the creation of Web Speech API, it has seen many Javascripts libraries created by developers who customized it for different purposes to enhance and maximise the use of API such as Annyang [3] and Artyom [4].

4 DATASETS

Before we present our system, we first provide an overview of the dataset used in our research. As the system targeted the 3rd Lifelog Search Challenge, we used the LSC'20 dataset [12]. The dataset consists of lifelog data from three sequential periods during the years 2015, 2016 and 2018 merged as one lifelog. In addition, the final dataset included images, visual concepts and metadata. There are approximately 190,800 images with every image represented by a minute ID. The visual concepts of data included text descriptions of detected scenes and concepts for each image. The metadata involves textual information showing time, physical activities such as *Speed*, biometrics such as *heart*, *calories*, locations such as *country name*, *geographic location* and others.

5 OVERVIEW OF VOXENTO

In this section, we present an overview of the system architecture with a detailed description of key components. There are three main components of Voxento as seen in figure 1: speech recognition, visual interface and underlying search engine. From a user's point of view, the system is web-based, but with a microphone (via a headset) for query interaction. Users will have two input methods by text or speech and an interface for visualisation and interaction and a voice-based response from the system. The system relies on the chrome browser to send the speech to the Google API interface and to also receive voice inputs. The system then sends the search task to the (indexed) search engine and retrieves the results in JSON format, ranked in terms of the most relevant text.

5.1 Voice-based Implementation

In order to provide speech recognition and synthesis functionality, the Google Web Speech API [24] was used, which facilitated Speech-to-Text and Text-to-Speech using the chrome browser. The Speech API was conferred to operate only in the English language (with English-British accent), though it can support other languages, and with additional APIs could also translate queries into English if needed. In order to detect parts-of-speech and grammar, we utilised the Java Speech Grammar Format, in the default setting, developed by Sun Microsystems. There are required configurations for the

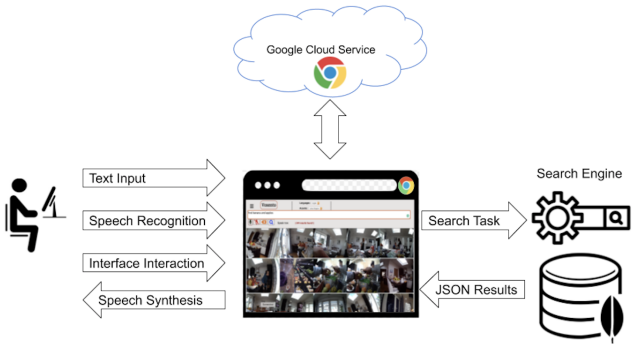

Figure 1: Overview Architecture

Google API, which were configured for continuous recognition, the provision of interim transcription (mid-speech) and the return of up to 3 maximum alternatives with scores. These parameters had the effect of allowing a user to continue speaking and see in real-time the updates to the result set as they continue to formulate a long query.

Speech recognition is only one element of a truly hands-free interface. Such a ubiquitous interface would also need to provide spoken output. In the current implementation of our Voxento prototype, the response is both spoken and visual with the majority of the post-query inspection via the visual interface. Initially, the voice feedback provides both status messages to the user (e.g. searching) and also communicates the number of results found. Additionally, image metadata is spoken back to the user when they select an image from the ranked list for detailed analysis.

With regard to the background noise which is a concern in our system, we noted that in LSC there is a noise caused by the number of participants in a space-limited room. Based on a phase of real-world testing, we found that the system will not be able to detect other voices than the main user unless they are sitting close to the microphone or speaking loudly. Consequently, using a headset device, we do not foresee any issue with using Voxento in the real-world environment.

5.1.1 Interactive Commands. In addition to simply speaking a query, Voxento supports system control via a number of commands, exemplified by the following three basic commands:

- **stop recording**: This allows users to stop recording and do any alteration on the text before they click submit manually, for example: *find a car and stop recording*. The system will stop recording and the final result will be: *find a car*. Users can update the query, for example by typing: *find a red car*.
- **submit**: Users can submit search tasks quickly and directly without clicking any button. An example would be: *find an apple and submit*. The system removes the words "and submit" before sending the query to the search engine and returning a ranked list.
- **reset**: This helps users when something goes wrong in the recording or when they want to change the query. They can say the **reset** command and the value of the search field will be erased.

It is important to note that the Google web speech API is configured to provide suggestions based on context and in our system, expressing commands in the wrong sequence or not in context, sometimes this affects accuracy although this did not occur often during testing. To illustrate this and to help users to avoid similar situations, we highlight the simple query: *find a car* followed by a **reset** command. The system can recognise this as: *find a car red*. Although the grammar is incorrect, the context seems fine as it appears to mean *a red car*. Thus, the suggestion is to use "and" and the query will be "find a car ...and reset". The dots in the query means if there is a pause in the middle, it will not affect the overall request.

5.2 Interface Design

The system has a clear and intuitive visual interface, as shown in figure 2. In this prototype, we have reduced the requirement for a user to use point-and-click interaction to a minimum (e.g. scrolling and selection of potentially relevant content from the ranked list). Consequently, the query panel is small (top of the screen) while the visual interface occupies the majority of the screen. The query panel facilitates the selection of language and accent, should they need to be changed from the default. In this implementation, we are not interested in changing language, though this could be a useful addition for a system developed for widespread use by novices. Below these configuration options is a large text box, which displays the output of the speech-to-text API and as such shows the current query text. This text can be manually modified with a keyboard if the user considers it necessary. Below this text-box are four buttons: start and stop recording, reset and submit button. Also, there are notices providing system status and the size of the ranked list. In order to start a voice interaction, the start recording button must be clicked to begin a session. The stop button can be used instead of voice command and the search button is also a convenient alternative to the automatic voice query submission.

Below this query panel is a simple visual ranked list of lifelog images. Should it be considered necessary, clicking on an image zooms displays it large in a floating panel.

5.3 Search Engine

The focus of this work is on the interaction components and as a result, we implemented a standard retrieval system using MongoDB and Mongoose API to support the search engine. We discuss how the search function works and the necessary changes made to perform a good search. Firstly, the dataset has two indexes: the first is minute_id, which includes date and time together, indexed in ascending order; the second is where all textual fields are indexed within the collection, for example, attributes and concepts. We use a free-text ranking engine as standard in MongoDB, which applies stemming and stop word removal. The maximum number of results returned for a query is configured to be 1,000. The system will show the first n results with the remainder coming slowly. Moreover, the system displays the top-N images in decreasing rank order in the visual panel. We select specific attributes in the results to be returned for minimising the effort of processing and focusing on the important fields.

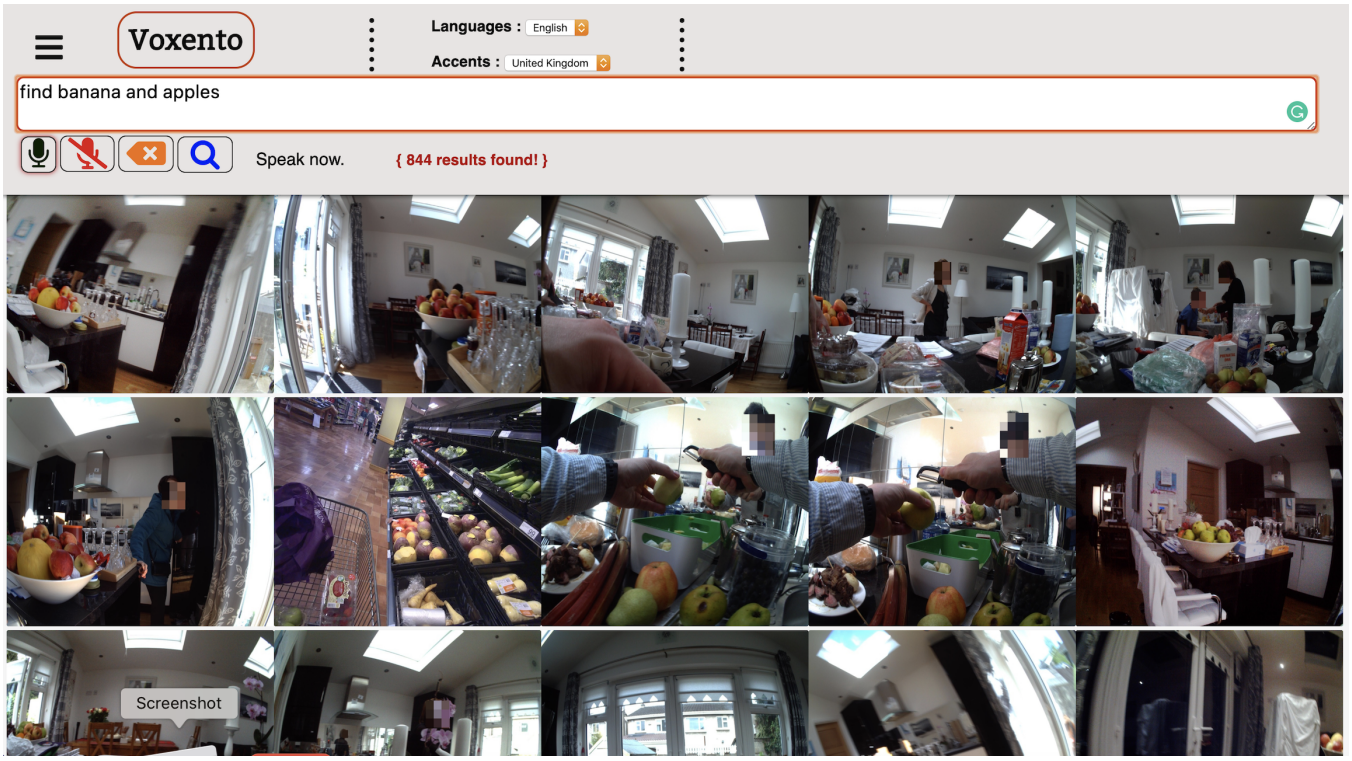

Figure 2: Visual Interface: Snapshot Overview

One challenge we face is that since Voxento uses speech input, which could have errors, this may affect the accuracy of results. However, for this prototype, we assume that such errors do not have a large impact on the retrieval process. For the weighting of terms in the ranking engine, we use the default weighting scheme in this first version of Voxento as we need to have a simple, clear, straightforward full-text ranking engine. See figure 3 for a sample image displayed in our system based on the query: *find apples*.

The final aspect of Voxento to discuss is the potential for application of NLP to the query strings. Since users will utter longer queries and more conversational when using Voxento, it becomes possible to extract parts of the speech to use them as a filter. For example, a query like: *I was in Russia, attending a conference, it was in 2018*, the system will extract "2018 year" as a faceted filter over the dataset, and avoids slow interface widget manipulation that would be necessary for a conventional *point-and-click* interface.

Figure 3: A sample image from LSC'20 with associated metadata: "2018-05-15_07:28", "home_office", "apple", "enclosed area", "glass".

6 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced a prototype interactive lifelog retrieval system called Voxento which uses speech recognition as an interaction mechanism. We highlighted the interactive features of Voxento, which is the first prototype of a voice-interaction search engine for lifelogs designed to be used in a pervasive and ubiquitous computing platform, such as google glass. In this first prototype, the voice interaction operated seamlessly with a conventional visual ranked list on a desktop computer. We described how the voice interaction works, the interface layout and how the search engine indexes and ranks the LSC'20 dataset.

There are a few limitations to this first version of the prototype. Firstly, and most obviously, it still requires a conventional desktop interface and mouse interaction. This will be replaced in

later versions with a fully wearable platform. Secondly, the underlying search engine is very simple, based on a conventional text retrieval approach and integrates no optimisation elements such as query/index expansion, embedding, enhanced metadata, etc. Such considerations were not the focus of this work. Finally, the commands for the system are still very simple and based on a standalone query. Ideally, such a query engine will support dynamic query modification and enhancements (e.g. 'no, not apples, I mean oranges…', or 'make it a Thursday…in the evening…after football practice'). Such a dynamic and persistent query engine will be better suited to the challenges of the LSC competition.

In terms of future work and improvements, the query processing engine will move from standalone to dynamic persistent querying, resulting in a more conversational interface for the LSC'21 prototype. Additionally, the effectiveness of the underlying search engine needs to be improved, which will be based on implementing the state-of-the-art approaches that are shown to be effective in the current LSC challenge. Finally, we intend to migrate from the desktop environment to a truly pervasive computing environment by developing the retrieval engine for an AR or VR platform in the next iteration.

ACKNOWLEDGMENTS

We acknowledge the support of Science Foundation Ireland and the Insight Centre for Data Analytics through the grant number SFI/12/RC/2289-P2 and the Ministry of Education in Saudi Arabia for sponsoring the PhD research of the primary author.

REFERENCES

[1] Julius Adorf. 2013. *Web Speech API.* Technical Report. 1–11 pages. http://home.in.tum.de/{~}adorf/pub/web-speech-api.pdf

[2] Eva Inaiyah Agustin, Riky Tri Yunardi, and Aji Akbar Firdaus. 2019. Voice recognition system for controlling electrical appliances in smart hospital room. *Telkomnika (Telecommunication Computing Electronics and Control)* 17, 2 (2019), 965–972. https://doi.org/10.12928/TELKOMNIKA.V17I2.11781

[3] Annyang. 2016. *annyang! SpeechRecognition that just works.* Retrieved March 10, 2020 from https://www.talater.com/annyang/

[4] Artyom. 2016. *Voice commands and speech synthesis made easy.* Retrieved March 10, 2020 from https://sdkcarlos.github.io/sites/artyom.html

[5] Clark Boyd. 2018. *Speech Recognition Technology: The Past, Present, And Future.* Retrieved March 10, 2020 from https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf

[6] Georgina Browne, Emma Berry, Narinder Kapur, Steve Hodges, Gavin Smyth, Peter Watson, and Ken Wood. 2011. SenseCam improves memory for recent events and quality of life in a patient with memory retrieval difficulties. *Memory* 19, 7 (2011), 713–722.

[7] Duc Tien Dang-Nguyen, Klaus Schoeffmann, and Wolfgang Hurst. 2018. LSE2018 panel - Challenges of lifelog search and access. In *LSC 2018 - Proceedings of the 2018 ACM Workshop on the Lifelog Search Challenge, co-located with ICMR 2018.* 1–2. https://doi.org/10.1145/3210539.3210540

[8] Martin Dodge and Rob Kitchin. 2007. 'Outlines of a World Coming into Existence': Pervasive Computing and the Ethics of Forgetting. *Environment and Planning B: Planning and Design* 34, 3 (2007), 431–445. https://doi.org/10.1068/b32041t arXiv:https://doi.org/10.1068/b32041t

[9] Aaron Duane, Cathal Gurrin, and Wolfgang Huerst. 2018. Virtual reality lifelog explorer :Lifelog search challenge at ACM ICMR 2018. In *LSC 2018 - Proceedings of the 2018 ACM Workshop on the Lifelog Search Challenge, co-located with ICMR 2018.* 20–23. https://doi.org/10.1145/3210539.3210544

[10] Sadaoki Furui. 1995. Speech recognition-past, present, and future. *NTT review* 7, 2 (1995), 13.

[11] April Glaser. 2017. *Google'S Ability To Understand Language Is Nearly Equivalent To Humans.* Retrieved March 10, 2020 from https://www.vox.com/2017/5/31/15720118/google-understand-language-speech-equivalent-humans-code-conference-mary-meeker

[12] Cathal Gurrin, Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Björn Þór Jónsson, Jakub Lokoč, Wolfgang Hurst, Minh-Triet Tran, and Klaus Schoeffmann. 2020. An Introduction to the Third Annual Lifelog Search Challenge, LSC'20. In *ICMR '20, The 2020 International Conference on Multimedia Retrieval.* ACM, Dublin, Ireland.

[13] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Duc-Tien Dang-Nguyen, Michael Riegler, Luca Piras, Minh-Triet Tran, Jakub Lokoč, and Wolfgang Hürst. 2019. [Invited papers] Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC2018). *ITE Transactions on Media Technology and Applications* 7, 2 (2019), 46–59. https://doi.org/10.3169/mta.7.46

[14] Cathal Gurrin, Alan F Smeaton, and Aiden R Doherty. 2014. *LifeLogging: Personal Big Data Foundations and Trends R in Information Retrieval.* Vol. 8. 1–107 pages. https://doi.org/10.1561/1500000033

[15] Wolfgang Hürst, Kevin Ouwehand, Marijn Mengerink, Aaron Duane, and Cathal Gurrin. 2018. Geospatial Access to Lifelogging Photos in Virtual Reality. 33–37. https://doi.org/10.1145/3210539.3210547

[16] H. J. Jeong, S. Ye, J. Lim, I. You, W. Hyun, and H. K. Song. 2013. A Remote Computer Control System Using Speech Recognition Technologies of Mobile Devices. In *2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing.* 595–600. https://doi.org/10.1109/IMIS.2013.105

[17] B H Juang and Lawrence R Rabiner. 2004. Automatic Speech Recognition – A Brief History of the Technology Development. *Elsevier Encyclopedia of Language and Linguistics* 50, 2 (2004), 637–655.

[18] Jochen Meyer, Steven Simske, Katie A Siek, Cathal G Gurrin, and Hermie Hermens. 2014. Beyond Quantified Self: Data for Wellbeing. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14).* Association for Computing Machinery, New York, NY, USA, 95–98. https://doi.org/10.1145/2559206.2560469

[19] Isadora Nguyen Van Khan, Pranita Shrestha, Min Zhang, Yiqun Liu, and Shaoping Ma. 2019. A Two-Level Lifelog Search Engine at the LSC 2019. 19–23. https://doi.org/10.1145/3326460.3329158

[20] Kieron O'Hara, Mischa M. Tuffield, and Nigel Shadbolt. 2009. Lifelogging: Privacy and Empowerment with Memories for Life. *Identity in the Information Society* 1, 1 (December 2009), 155–172. https://eprints.soton.ac.uk/267123/

[21] Zhengwei Qiu, Cathal Gurrin, and Alan F. Smeaton. 2016. Evaluating Access Mechanisms for Multimodal Representations of Lifelogs. In *MultiMedia Modeling*, Qi Tian, Nicu Sebe, Guo-Jun Qi, Benoit Huet, Richang Hong, and Xueliang Liu (Eds.). Springer International Publishing, Cham, 574–585.

[22] Luca Rossetto, Ralph Gasser, Silvan Heller, Mahnaz Amiri Parian, and Heiko Schuldt. 2019. Retrieval of Structured and Unstructured Data with vitrivr. In *Proceedings of the ACM Workshop on Lifelog Search Challenge - LSC '19.* 27–31. https://doi.org/10.1145/3326460.3329160

[23] Stefan Selke and Philipp Klose (Eds.). 2016. *Lifelogging : digital self-tracking and lifelogging - between disruptive technology and cultural transformation.* Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-13137-1

[24] Glen Shires. 2013. *Voice Driven Web Apps: Introduction to the Web Speech API.* Retrieved March 10, 2020 from https://developers.google.com/web/updates/2013/01/Voice-Driven-Web-Apps-Introduction-to-the-Web-Speech-API

[25] Qianling Zhou, Di Wang, Cliona Ni Mhurchu, Cathal Gurrin, Jiang Zhou, Yu Cheng, and Haijun Wang. 2019. The use of wearable cameras in assessing children's dietary intake and behaviours in China. *Appetite* 139 (2019), 1–7. https://doi.org/10.1016/j.appet.2019.03.032