IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 January 2023, accepted 3 February 2023, date of publication 23 February 2023, date of current version 30 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3248284

RESEARCH ARTICLE

Comparing Interactive Retrieval Approaches at the Lifelog Search Challenge 2021

LY-DUYEN TRAN[1], MANH-DUY NGUYEN[1], DUC-TIEN DANG-NGUYEN[2,3], (Member, IEEE), SILVAN HELLER[4], (Student Member, IEEE), FLORIAN SPIESS[4], JAKUB LOKOČ[5], LADISLAV PEŠKA[5], THAO-NHU NGUYEN[1], OMAR SHAHBAZ KHAN[6], AARON DUANE[6], BJÖRN ÞÓR JÓNSSON[6], LUCA ROSSETTO[7], AN-ZI YEN[8], AHMED ALATEEQ[1], NAUSHAD ALAM[1], MINH-TRIET TRAN[9], (Member, IEEE), GRAHAM HEALY[1], KLAUS SCHOEFFMANN[10], AND CATHAL GURRIN[1]

[1]Adapt Centre, the School of Computing, Dublin City University, Dublin 9, D09 DXA0 Ireland
[2]Department of Information Science and Media Studies, University of Bergen, 5007 Bergen, Norway
[3]School of Economics, Innovation and Technology, Kristiania University College, 0107 Oslo, Norway
[4]Databases and Information Systems Group, University of Basel, 4001 Basel, Switzerland
[5]Faculty of Mathematics And Physics, Charles University, 11000 Prague, Czech Republic
[6]Department of Computer Science, IT University of Copenhagen, 2300 Copenhagen, Denmark
[7]Department of Informatics, University of Zurich, 8006 Zürich, Switzerland
[8]Department of Computer Science, National Yang Ming Chiao Tung University, Taipei 112, Taiwan
[9]Faculty of Information Technology, University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[10]Institute of Information Technology, Klagenfurt University, 9020 Klagenfurt am Wörthersee, Austria

Corresponding author: Ly-Duyen Tran (ly.tran2@mail.dcu.ie)

This work was supported in part by the Science Foundation Ireland Centre for Research Training under Grant 18/CRT/6223 and Grant 18/CRT/6224, in part by the Insight Science Foundation Ireland (SFI) Research Centre for Data Analytics under Grant SFI/12/RC/2289_P2, in part by the Czech Science Foundation (GAČR) under Project 19-22071Y, in part by the ADAPT—Centre for Digital Content Technology under Grant SFI/13/RC/2106_P2, in part by the Swiss National Science Foundation through the Project "Participatory Knowledge Practices in Analogue and Digital Image Archives" under Contract CRSII5_193788, and in part by Charles University under Grant SVV-260588.

ABSTRACT The Lifelog Search Challenge (LSC) is an interactive benchmarking evaluation workshop for lifelog retrieval systems. The challenge was first organised in 2018 aiming to find the system that can quickly retrieve relevant lifelog images for a given semantic query. This paper provides an analysis of the performance of all 17 systems participating in the 4th LSC workshop held at the 2021 Annual ACM International Conference on Multimedia Retrieval (ICMR). LSC'21 was the largest effort at comparing different approaches to interactive lifelog retrieval systems seen thus far. Findings from the challenge suggest that many different interactive factors contribute to the success (or otherwise) of participating teams. In this paper, we provide an overview of the LSC'21 challenge, introduce each team's approach and explore these factors in depth and offer clues on how to develop a high-performing interactive lifelog search engine.

INDEX TERMS Lifelog, information retrieval, multimodal, analytics.

I. INTRODUCTION TO THE LIFELOG SEARCH CHALLENGE

In recent years, the increase in volume of personal multimedia data from wearable computing devices has created a need for reimagining how large volumes of personal data can be organised. Specifically, we note the ubiquity of self-quantification devices [1] and the increasing prevalence of point-of-view (PoV) cameras. Consequently, the annual Lifelog Search Challenge (LSC) was started in 2018 to provide a collaborative and open forum for researchers to compare approaches to interactive retrieval from large personal data archives, with a specific focus on lifelog data from wearable sensors, such as chest-mounted cameras, biometric wearables, and location trackers. The LSC challenge has been instantiated as a workshop at the ACM ICMR conference, and the 4th Lifelog Search Challenge (LSC'21) took place during ACM

The associate editor coordinating the review of this manuscript and approving it for publication was Fu Lee Wang.

ICMR'21, in November 2021. It has been the largest LSC workshop to date with 17 participating teams. Due to the pandemic, the search challenge was organised as a virtual workshop, where participants and organisers were connected in a video conferencing session. Each participating system took part in a synchronised competition which required the interactive processing of a wide range of information needs and the performance of each system was calculated in real-time and displayed on a shared scoreboard. The best performing system was identified at the end of the search challenge, which lasted about two hours.

At LSC'21, each of the participating teams brought a unique and customised search engine to the challenge. In this paper, we introduce the LSC challenge, describe all competing systems, and highlight the techniques and components that are employed in state-of-the-art interactive lifelog retrieval systems. We conclude by suggesting how future interactive lifelog retrieval challenges can be improved. To the best of our knowledge, LSC'21 is the largest iteration of the challenge at the time of writing. Since the comparative analysis of LSC'18 [2], many approaches have been proposed and refined for the task of lifelog retrieval. A thorough investigation is important to define the future of this research area. The contributions of this work are, firstly, the introduction to a novel collaborative retrieval challenge, the review of the state-of-the-art approaches, and a detailed analysis of system performance leading to clues as to how to develop a next-generation of interactive lifelog retrieval system.

II. THE LIFELOG SEARCH CHALLENGE

The Lifelog Search Challenge is a participation workshop in which teams compete with each other to develop the leading interactive lifelog retrieval tool. The aim of the challenge is to provide an open and collaborative environment for participants to benchmark the performance of their interactive retrieval systems and learn from the performance of their systems and competitor systems. It is anticipated that the open, shared, metrics-driven evaluation will lead to an increase in the performance of all retrieval systems for lifelogs.

The challenge is organised as a synchronous, live competition in which a large number of tasks (information needs) are presented in sequence to the participants who must solve each task and submit the correct answer to a host server, which calculates scores and displays system performance on a shared scoreboard. Each task is represented as a form of known-item search with a unique relevant lifelog event. Given a time-limit of five minutes, tasks are presented to all participants in a synchronous manner, meaning that all participants see the same task at the same time and see their position on the scoreboard in real-time.

We will now describe the dataset and tasks used for the challenge. At the LSC workshop, datasets are usually used for two sequential years before being replaced by a larger dataset.

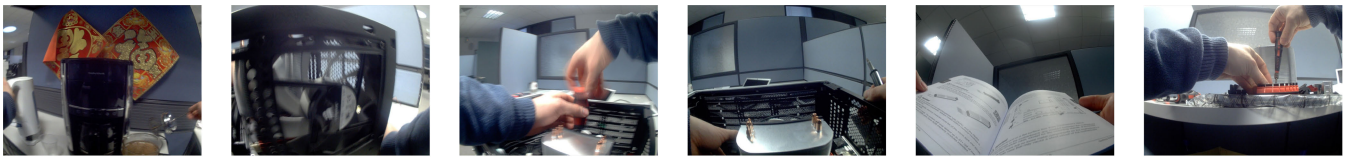
A. LSC DATASET

To support the comparative challenge, a four-month lifelog dataset from one individual lifelogger was distributed to all participants a number of months before the challenge took place. This dataset had also been previously employed in the LSC'20 workshop, though the version used in LSC'21 was slightly smaller in size, having had 8,126 images removed for data governance reasons. The dataset had been constructed by merging parts of the three NTCIR Lifelog datasets from 2016 [3], 2017 [4], and 2019 [5], with images from 2015, 2016 and 2018, respectively. It consisted of an image dataset of 183,299 wearable camera images at a resolution of 1024 × 768 (37.35GB). These images were captured using either an OMG Autographer or Narrative Clip wearable camera, typically at a rate of around one every 30 seconds during waking hours. These devices are worn clipped onto clothing or on a lanyard around the neck, facilitating the capture of a first-person view of the life experiences and activities of the lifelogger. Prior to release, all images were anonymised, which means that faces and most readable text on screens were redacted in a manual or semi-manual process. Additionally, there was an associated metadata file consisting of timestamps (on a minute-by-minute basis, physical activities, detailed minute-by-minute biometrics (for all years except 2015), locations of the individual, and for each image, a list of visual concepts extracted from the non-redacted version of the image dataset, which includes bounding boxes for objects. Example images from the collection are shown in Figure 1 (below). It is worth noting that a unique aspect of the LSC challenge is that the lifelog dataset includes metadata captured 24 × 7 and point-of-view wearable camera images captured all day, during waking hours.

B. TASKS & RELEVANCE JUDGEMENTS

For the LSC'21 benchmarking workshop, 24 tasks were prepared with a textual query and a manually generated ground truth. Each task represented an information need and most were generated by the lifelogger who created the collection. The tasks were selected to represent important life activities that only occurred once (or very few times) within the dataset, and as such represented a form of known-item search. Tasks were formed with a single information need in mind, but were constructed in a temporally advancing manner, with each task being composed of six increasingly detailed task descriptors, which were revealed at six points (0, 30, 60, 90, 120, and 150 seconds). The final (detailed) subtask remained on screen for 150 seconds, meaning that each task had an allocated time of five minutes. An example of LSC'21 tasks is detailed in Table 1 and some images from the ground truth are shown in Figure 1.

The participating teams were required, for each task, to find any relevant image and submit it to a host server [6]. The host server maintained a countdown clock and actively evaluated submissions against the ground truth.



Numbers of images: 92
Time: 07:24AM - 09:27AM
Date: 13/03/2015 (Friday)
Semantic Name: Dublin City University (DCU)

FIGURE 1. Some lifelog images as ground truth from task I in Table1.

TABLE 1. Task 1 with its temporally advancing descriptors, which were revealed at 30-second intervals. After 150 seconds, the full description is shown for another 150 seconds until the end of the task.

Time	Text
0s	I was building a computer alone in the early morning on a Friday...
30s	I was building a computer alone in the early morning on a Friday at a desk...
60s	I was building a computer alone in the early morning on a Friday at a desk with a blue background...
90s	I was building a computer alone in the early morning on a Friday at a desk with a blue background. Sometimes I needed to refer to the manual...
120s	I was building a computer alone in the early morning on a Friday at a desk with a blue background. Sometimes I needed to refer to the manual. I remember some Chinese posters on the desk background...
150s	I was building a computer alone in the early morning on a Friday at a desk with a blue background. Sometimes I needed to refer to the manual. I remember some Chinese posters on the desk background. I was in Dublin City University in 2015.

C. SCORING

For each task, its score depends on the time of the first correct submission and the number of previous incorrect attempts by the team. The score ranged from 0 to 100, with an overall normalised score continuously displayed for each team, aggregating the scores for the tasks up until that point. For a given task, the score is calculated as follows, and is identical to the KIS scoring function at the VBS challenge [7]. Given a linearly decreasing function f_{TS} based on search time, the time of correct submission t and the number of wrong submissions ws, the score (TS) for a given task for a single team is as shown in equation (1) below:

$$f_{TS}(t, ws) = max(0, 50 + 50 \cdot f_{TS}(t) - 10 \cdot ws) \qquad (1)$$

f_{TS} therefore results in at least 50 points for a correct submission if no incorrect submission was previously made, and penalises each wrong submission by 10 points. For more details, see [7].

III. PARTICIPANT TEAM OVERVIEWS

In LSC'21, 17 teams took part in the benchmarking challenge. Each team had had access to the full dataset to process and index the data a number of months prior to the challenge. In this section, we briefly introduce each system. Readers are encouraged to read the individual participant papers describing each system from the LSC'21 proceedings [8].

The **MySCéal** system [9] was a second generation retrieval system which first participated at LSC'20 [10]. This system was designed to support novice users who would not be familiar with interactive lifelog retrieval systems. Therefore, the tool was designed with a straightforward user interface based on textual queries. Additionally, it integrated a novel scoring function measuring the similarity between semantic textual inputs and the annotations of images, which were visual concepts such as objects or OCR detected in images, to find relevant images. Furthermore, the interface of the system contained a map visualising the GPS data which could be used as a filtering mechanism according to geographic bounds.

The **SomHunter+** [11] team participated with an extended version of the SOMHunter tool from previous VBS/LSC events, which relied on a new text search model based on CLIP [12] (developed by Open AI) that proved to be effective on a large variety of different datasets. The system utilized also W2VV++ model [13] for similarity search and localized queries. For results of a query, the tool supported a grid based visualization as a ranked list or as a relevance-aware self-organizing map. In addition, the tool supported temporal and localized queries, allowing users to express both temporal dependencies and location of searched items. However, the users found that the structure of the tasks at the competition led mostly to pure query-and-scroll search strategy, which was sufficient for most of the solved tasks. Especially for tasks with known displayed text in the searched image, the CLIP model demonstrated notable effectiveness in OCR.

LifeSeeker [14] is the third version of the concept-based retrieval system firstly introduced in LSC'19 [15]. The key idea behind the search mechanism is to leverage the image's annotations, including visual concepts, text and additional metadata, in order to retrieve the desire moments by means of two different models, elastic search and weighted bag-of-words. By extracting more visual objects using the Microsoft vision API, refining GPS data information and defining semantic location names, indexed concept based knowledge is enriched leading to an increase in searching speed and accuracy.

Voxento [16] is an interactive voice-based retrieval system for lifelogs which has participated since LSC'20 [17] and provides a spoken interface to the lifelog dataset, which facilitates users to interact with a personal lifelog using a range of vocal commands and interactions. The system relies on Google web speech API for speech recognition and synthesis. The user has the choice to use a mouse and keyboard or voice interaction at any stage during the interaction with the interface. Voxento at LSC'21 employed the backend API provided by [18], which uses the CLIP model for text-to-image search based, with minor modifications to support some interaction requirements of Voxento.

The **CVHunter** [19] system was tested as a ''rapid-development'' based application (in WPF .NET) created in a short time period before the competition. The application used the same metadata as SomHunter+ and provided basic browsing functions like ranked set scrolling, day summary browsing, and query by example image search. According to the results of CVHunter, this experiment demonstrated that with already available state-of-the-art content-based features, it is possible to relatively quickly design/implement a simple yet competitive system for the evaluation benchmark.

Memento [18] is a prototype lifelog search engine which participated for the first time in LSC'21. The system implemented image-text embeddings derived using the CLIP [12] model. The system also accepts queries in natural language allowing the user to specify visually complex scenarios quite easily. Memento further supports searching for events with a temporal context allowing users to search for a target event in the context of a past or future event. The user interface of the system includes a primary search interface to initiate search and view results, a starring feature to tag probable images during an ongoing search, and a data filtering component to support faceted filtering while supporting data visualization of the ranked results to aid better decision-making during the challenge.

FIRST [20] is an interactive retrieval system that supports multiple modalities for interaction and query processing, including textual querying, query by meta-data (including date, time, and location) Additionally FIRST also utilises extract scene text, entities, activities, places to enrich the provided meta-data for each image. Text and visual information matching is based on joint embedding model. Scene clustering is based on visual and location information, facilitating two types of content clustering. The system also integrates a flexible timeline to shrink or expand the time interval of interest and query expansion with visual examples for visual similarity-based ranking.

NTU-ILRS (LifeConcept) [21] is an interactive visual lifelog retrieval system that also utilises word embeddings to reduce the semantic gap between textual queries and the images. Additionally, the relation graphs within images are extracted by employing the Multi-Level Scene Description Network [22] since the queries may describe relationships among objects. As a result, word-level and sentence-level embeddings are encoded into the framework. To further confirm the information need of the user, the system provides a list of relevant concepts of the query terms by consulting the recommendation from ConceptNet[1] for user selection. Finally, a ranking mechanism is employed that is based on the BM25 scoring function to search the relevant images and display them to the user.

lifeXplore [23] is a visual lifelog retrieval system built for temporal structuring and filtering of lifelog data. The system consists of a MongoDB[2] database and a Node.js[3] back-end that interacts with an Angular front-end that provides rich search and filtering features. The interface allows to filter data by time-related groups, such as years, months, days, weekdays, day-times, etc., and to combine such filters with content-based search for contained objects, full-frame concepts, and recognized text. Results can be grouped by coherent days, further inspected by a meta-data viewers, or they can be used for content-based similarity search to explore other similar data items.

LifeMon [24] is a new prototype system aimed at exploring the suitability of MongoDB for lifelog storage and retrieval. The system maps the LSC metadata into two document schemas and provides a browser-based filtering interface based on the commonly used metadata, such as date and time, concepts, attributes and categories. Users can examine result images in detail and traverse their temporal context. With suitable indexing, the performance was sub-second for most tasks on a moderate laptop, indicating a potential for this architecture.

vitrivr [25] is an open-source[4] multimedia retrieval system [26] which has previously participated at the LSC workshop [27], [28]. It utilizes a dedicated database for multimedia features [29], [30] and a retrieval engine [31] supporting different query modalities and media types. Query-by-sketch, query-by-example, semantic concepts, and query-by-motion are some examples of supported retrieval models. For LSC'21, the system introduces an image stabilisation module to address the lifelogger's movements, as well as support for map-based queries.

vitrivr-VR [32] is a multimedia retrieval system in Virtual Reality which builds upon the vitrivr stack, utilizing the same retrieval engine and database. It has shown competitive performance at interactive competitions [33], [34]. It offers novel VR-based ways to present and interact with results, such as a cylindrical result presentation view and a multimedia drawer with which results of a single day can be quickly explored in VR.

Exquisitor [35] is a research prototype aimed at studying the role of interactive learning in large-scale multimedia analytics applications. The overall goal of Exquisitor is to evolve a semantic classifier, in cooperation with the user using relevance feedback, to capture the user's information

[1] http://conceptnet.io
[2] https://www.mongodb.com/
[3] https://nodejs.org/en/
[4] https://vitrivr.org

need [36]. To support rapid semantic model construction and collection exploration, Exquisitor provides a search capacity to identify positive examples, filtering to focus the scope of exploration, and timeline-bases exploration. Furthermore, to support tasks with a temporal component, Exquisitor allows building multiple classifiers and retrieving segments that satisfy both, optionally with a constraint on the temporal relationship. Due to a late issue with indexing semantic labels, which in turn led to incorrect mappings between images and their labels, the system performed worse than in previous competitions.

XQC [37] relies on the interactive learning engine of the Exquisitor system, but provides a cross-platform interface for collection exploration. The XQC system has most of the functionality of the Exquisitor system, aside from temporal operators, and therefore suffered the same index mapping issue as its parent system. The main goal of the XQC team was to investigate the feasibility of a mobile interface, however, and they took part using one Android-based mobile phone. The fact that XQC performed comparably with Exquisitor indicates that mobile interfaces have perhaps surprising potential.

PhotoCube [38] is a prototype for a multimedia analytics system, based on the Multidimensional Media Model (M^3, pronounced "emm-cube"). The M^3 model proposes to map a media collection to a multi-dimensional metadata space [39], supporting faceted exploration of these dimensions and mapping the result of the faceted filters to an exploration cube in 1–3 dimensions. The PhotoCube prototype consists of a media server, implemented in PostgreSQL, and a browser-based exploration client. The M^3 model is intended for collection exploration, rather than item-based media retrieval, and the current prototype does not support efficient query construction, so PhotoCube was expected to poorly match the tasks of the LSC competition.

ViRMA [40] is a virtual reality multimedia exploration prototype which shares the same back-end server as PhotoCube [38], with both systems utilising the Multidimensional Multimedia Model (M^3) [39], which supports the browsing of multimedia data objects by translating them from multidimensional media space to 3D space. This can then be easily mapped to 3D virtual space in ViRMA so that the user can navigate and browse the output from their filter queries in a direct and intuitive way. However, as was the case with PhotoCube, the prototype's features did not translate very effectively to the LSC competition as the system is tuned toward exploration, rather than search, and the LSC tasks are currently very search-focused. Due to this, the researchers intend on supporting more explicit exploration and browsing tasks in the future, such as assisting in the generation of search tasks for future LSC challenges.

LifeGraph [41] is an experimental Knowledge Graph-based lifelog retrieval system that participated to LSC for the second time in 2021. The underlying graph is built from instances of everyday objects that have been automatically detected in the lifelog images. The detected concepts are then linked to an external knowledge base[5] in order to enrich them with additional context, enabling the indirect retrieval of higher-level semantic concepts which cannot easily be detected directly. Querying is done by selecting an arbitrary number of graph nodes as start points. The graph is then traversed until a sufficiently large number of log entries is reached. The score of a result is inversely proportional to its graph distance from the start points. An embedding method originally developed for link prediction [42] is used to query for similar log entries based on previously retrieved ones. LifeGraph re-uses several components from the vitrivr stack. It shares the underlying database system [29] and uses modified versions of the retrieval engine as well as the user interface.

IV. A REVIEW OF APPLIED TECHNIQUES

The 17 systems just introduced utilise a wide variety of underlying technologies. We will highlight the main approaches taken by the participating systems in the following section and in table 2. These approaches can be seen as an indication of the state of the art components of a modern interactive lifelog search engine. Please note that the below summaries are not-exhaustive. Given space limitations, we have limited the number of systems that we highlight for each technical approach. The interested reader should review table 2 and find more details in the associated participant papers.

A. CONCEPT-BASED RETRIEVAL

Concept-based search is a conventional approach to retrieval from visual media archives that employs object detection techniques to associate visual concepts with an image. Each image is assigned a list of visual concepts which are then used in various techniques to match with the query. For the LSC'21 challenge, the concepts generated by Microsoft's Vision API,[6] utilised by MyScéal at LSC'20, were provided to every participating team before the live challenge. In addition, the teams may choose more computer vision models to attain the concepts.

All participating teams implemented some form of concept-based retrieval, whether as a source for semantic matching with textual queries, or as a form of faceted filtering. Some groups took a more novel approach to employing visual concepts. For example, MyScéal [10] uses aTF-IDF, a modified version of TF-IDF, which exploits the provided object annotations by incorporating the object's area into scoring. Considering that retrieving the images may require understanding the relative position of objects, NTU-ILRS incorporates the relation graph of objects with the visual concepts of images. Moreover, NTU-ILRS expands the query terms from the given queries by utilising ConceptNet to produce additional suggested terms to assist the user to generate suitable query terms. Other participants, such as FIRST,

[5] https://www.wikidata.org/

[6] https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

extract objects using Faster RCNN [43], EfficientDet [44], and their own object detectors for items that appear in daily life activities.

B. MULTIMODAL EMBEDDINGS

Embedding-based search has shown great promise to support effective content retrieval between modalities. In embedding-based search, both natural language queries and the images from the dataset are mapped into a common space, where their similarity is evaluated. Cosine similarity is a popular choice for comparison and is defined as:

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

where $\mathbf{q}$ is the encoded search query, and $\mathbf{c}$ is the encoded image.

Multimodal embedding models have gained significant attention in recent years. They are also frequently used in state-of-the-art video retrieval systems [7]. As for LSC'21, more than half of the teams apply this approach to their system. For instances, both vitrivr and vitrivr-VR utilise an approach similar to W2VV++ [13] originally developed and used in video retrieval [33], [34], [45]. In NTU-ILRS, the visual concepts provided by the object detection model from Microsoft Vision API are encoded with FastText [46] while the generated image captions and the user's queries are encoded with Sentence BERT [47]. The images are then retrieved using cosine similarity scores between query embeddings and image embeddings constructed by the image-to-text and image-to-caption models [48]. FIRST used their own Self-Attention-based Joint Embedding Model (SAJEM) [20] using Faster R-CNN Bottom-Up to encode images and RoBERTa to encode text description. Memento [18], SomHunter+ and CVHunter used image-text embeddings derived from the zero-shot CLIP model [12] and ranks the images by comparing the query vector with the image vectors on the basis of cosine similarity scores. One notable advantage of the CLIP model over, e.g., W2VV++ is the usage of sub-word encoding in the text embedding branch. This makes the model more robust against various nuances or typos in the text query and allows users to construct more natural descriptions of the searched image, rather than ''keyword-style'' descriptions. The impressive performance of CLIP has led to its adoption by teams in the following year [49].

C. OPTICAL CHARACTER RECOGNITION (OCR)

This year's challenge introduced a new trend in formulating search queries which relied on the presence of visible text in the target images. This had been hinted at in the previous year's competition, where some queries would clearly have benefited from the presence of OCR text extracted from the image content. One example of such queries is Task 14, presented in Table 1. With the first hint of '*I remember the TagHeuer advertisement for a watch*', a search for an image with the word 'TagHeuer' recognised by an OCR model would be sufficient to solve the task.

Various off-the-shelf OCR tools were employed by participants to enrich the searchable data with texts extracted from the visual content. One of such is Google's Cloud Vision API [7] that was used by MyScéal. However, it is not always obvious whether a hint in the query is searchable using OCR models. In such cases, embedding models, especially CLIP, as mentioned in the previous subsection, are seemingly a better choice as they support recognising both texts and semantic content in images. However, the OCR performance of CLIP is mixed [12] and might not be as optimised as the OCR models that are specially designed for this task.

D. TEMPORAL QUERY HANDLING

To address the temporal nature of many lifelog queries (e.g. the lifelogger did this, then that), some systems provide the functionality to search for multiple temporally ordered queries. MyScéal's supports three ordered queries, where the supplementary queries (before and after) are performed conditionally on the main one. That is, the main query is searched first. Then, for each result, a time filter is created (e.g. 2 hours later), which is used with the supplementary search query. Vitrivr and vitrivr-vr support an arbitrary number of temporally ordered queries, whose scores are fused together in a late fusion step [50], [51]. Exquisitor supports temporal query handling by building multiple semantic classifiers and merging the results of two classifiers based on temporal conditions [35]. Memento's [18] temporal search algorithm initially attempts to locate the target event and then in the subsequent stage re-ranks the initial list by searching for the past and future context provided by the user in a predefined search space. SomHunter+ and CVHunter both support querying via two temporally ordered queries, where the best match for the second query is considered from a fix-sized neighborhood of the first query's target [52]. For handling the temporal queries, LifeSeeker proposes Elastic Sequencing [53] to display next and previous images with respect to the current image in a sequence. The time distance between them can be manually adjusted to being temporally nearby or further apart.

E. RELEVANCE FEEDBACK

Relevance feedback is a commonly used technique to support a user's interaction with an interactive retrieval system. Exquisitor uses relevance feedback in which the user is presented a set of items they need to label either positive or negative. The feedback from the user is used to train a new or existing classifier, that is then used to retrieve a new set of suggestions to present the user with. SomHunter+ also supports iterative relevance feedback, but in contrast to Exquisitor, only positive examples are labeled,

[7] https://cloud.google.com/vision/docs/ocr

while negative ones are sampled from unlabeled items. SomHunter+ relies on a Bayesian-like updates of relevance probability governed by distances to positive and negative samples [54], [55]. Similar to the approach to video retrieval [56], vitrivr and vitrivr-vr support simple more-like-this queries using MobileNet V1[8] [57] NTU-ILRS, implements something similar in that their system asks the user to provide more details of their desired images, such as location, time, and so on, to filter out images not related to the queries.

F. VISUAL SIMILARITY

Image similarity can be considered to be a form of relevance feedback for image and video data; it is commonly used as a form of alternative search in which the user can request images that are visually similar to a specific image that they have already located. For image similarity, MyScéal combined features from a pretrained VGG16 model [58] with visual local features [59], [60], [61]. Likewise, FIRST and SomHunter+ used ResNet152 and W2VV++ [13] features respectively to calculate the similarities. Meanwhile, LifeSeeker utilises the Bag-of-Visual-Words model to transform visual images into numeric vector representations before implementing the K-means algorithm to cluster those visually similar images into groups of images.

Visual similarities can also be used in other ways. MyScéal joined temporally adjacent images that are visually similar to reduce visual clusters. In addition, FIRST exploited similarity scores between images to form clusters and display search results.

G. LOCATION VISUALISATION

Many groups facilitated map-based search on a map annotated with important semantic locations. In MyScéal, the search results are clustered in the map section along with the location names inferred by the query parser. The user can also draw a rectangle on the map to narrow the search space down to only the moments that happened inside that area. FIRST also visualises photo clusters based on geolocation on a map. Furthermore, vitrivr offered a map-based query interface using leaflet[9] and the leaflet-geosearch package[10] to query for arbitrary locations.

H. NOVEL INTERACTION

Most of the systems that participated this year utilised a desktop-based interface. However, some systems experimented on a Virtual Reality (VR) interface such as vitrivr-VR, PhotoCube, and ViRMA. Another team, XQC, employed the functionality of the Exquisitor system on a mobile interface, specifically Android-based. Moreover, both Voxento and vitrivr-VR used speech recognition to assist a user to formualte a query.

[8]https://tfhub.dev/google/imagenet/mobilenet_v1_050_192/quantops/feature_vector/3

[9]https://leafletjs.com/

[10]https://github.com/smeijer/leaflet-geosearch

TABLE 2. Selected approaches used by participating systems. For each system, a reference to the paper describing the method is given. Very common techniques, such as search/filter by visual concept and filter using the provided metadata (e.g. time, location), are not included in this table, since most systems implement some form of both.

	concepts search	embedding	OCR	temporal query	relevance feedback	visual similarity	location visualisation	novel interaction
MyScéal [9]	✓		✓	✓		✓	✓	
SomHumter+ [11]		✓		✓	✓	✓		
LifeSeeker [14]	✓		✓	✓		✓		
Voxento [16]		✓						✓
CVHunter		✓		✓		✓		
Memento [18]		✓		✓				
FIRST [20]	✓	✓				✓	✓	
NTU-ILRS [21]	✓	✓		✓	✓			
lifeXplore [23]								
LifeMon [24]	✓							
vitrivr [25]	✓	✓		✓	✓		✓	
vitrivr-VR [32]	✓	✓			✓			✓
XQC [36]	✓				✓			✓
Exquisitor [36]	✓			✓	✓			
PhotoCube [39]	✓							✓
ViRMA [39]	✓							✓
LifeGraph [41]	✓	✓						

V. PERFORMANCE ANALYSIS OF PARTICIPANTS

The LSC benchmark provides a single numerical score that reflects the relative performance of each system in the challenge, with the top system given a score of 100 and all other systems being scored in relation to this.

A. OVERALL SCORE

Table 3 illustrates the number of tasks that each system found the correct answers as well as its final normalized score. It can be seen from the table that although MyScéal achieved the highest score, Lifeseeker was the team that solved the most tasks compared to other participants. This is because the score was calculated not only based on the number of tasks solved by a team but also its accuracy in their submissions to find the answer of the tasks. Additionally, the search time for each task also plays a critical role in the scoring.

All lifelog retrieval systems participated managed to get at least 3 correct answers out of 23 tasks in total, however, there was no team able to solve all given queries. While there was a ranking of all teams, it is worth noting that there was actually very little difference between the top 3 teams, with a difference of 1 in the number of solved tasks and 3 points in the final score. MyScéal achieved the top performance and acquired the first place although the number of tasks they solved was the same with SomHunter+ and one task less than Lifeseeker. Despite the first time attending the LSC, CVHunter and Memento achieved a high performance with 15 and 16 solved tasks accordingly. Experience suggests that in the second and subsequent years, that participating teams achieve a high ranking as they refine and enhance an existing system, rather than create a new one for subsequent years.

TABLE 3. The overall scores of 17 participating teams with the number of tasks that they successfully solved.

	MySCéal	SomHunter+	LifeSeeker	Voxento	CVHunter	Memento	FIRST	NTU-ILRS	lifeXplore	LifeMon	vitrivr	vitrivr-VR	XQC	Exquisitor	PhotoCube	ViRMA	LifeGraph
Solved tasks	19	19	**20**	18	15	16	12	12	12	10	8	6	5	5	5	4	3
Score	**100**	97.6	97	91.4	77.3	77.2	55.5	48.3	47.4	38.7	32.5	27.7	21.5	21.2	19.4	16.0	11.5

B. NUMBER OF CORRECT/WRONG SUBMISSIONS

For a deeper understanding in performance of participants, Figure 2 depicts the number of correct and wrong submissions of each team made across all tasks in LSC'21. The blue bars indicate the number of correct answers (or solved tasks as shown in Table 3), whilst the orange bars illustrate the number of incorrect answers. VIRMA and vitrivr are the teams submitted the most with 43 and 36 submissions, respectively. They also had the most wrong answers compared to others which lowered their scores. The accuracy of submission is also a key factor in evaluating the scores of systems. This is one of the reasons why MySCéal recieved a higher final score than SomHunter+ and LifeSeeker although MySCéal had less correct answers than the latter teams. The same is true for XQC, Exquisitor and PhotoCube which all obtained 5 solved tasks but PhotoCube had the least score with more incorrect submissions than the other 2 teams.

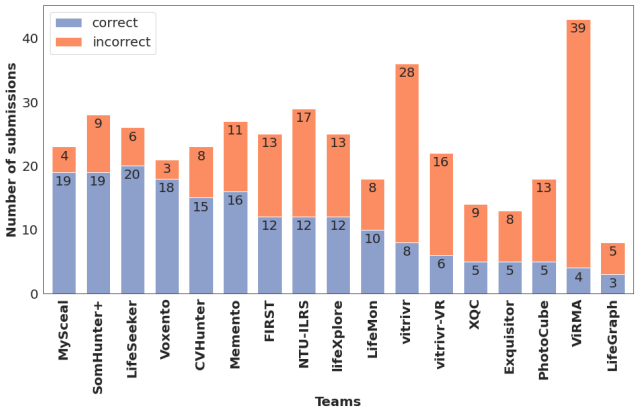


FIGURE 2. Number of correct/wrong submissions per team across all tasks.

Figure 3 shows the precision and recall of the submissions from each team. There is a clear gap between the top 4 ranked teams and the rest. which are MySCéal, SomHunter+, LifeSeeker, and Voxento. Indeed, given the subtleties of the scoring mechanism, all of these four teams could be considered the top performing teams with equivalent performance. Additionally, given the fact that a major factor in the performance of an interactive system is the skill and expertise of the searcher, it becomes even more difficult to differentiate between the performance of the top ranked systems.

With the least number of incorrect submissions (see Figure 2), Voxento managed to achieve the highest precision among all participants with 18 correct answers out of 21 submissions. In contrast, the recall of the top performaing systems shows a small difference. LifeSeeker obtained the highest recall when they could solved 20/23 queries which is slightly better than MySCéal, SomHunter+ and Voxento. It is interesting to point out that SomHunter+ got higher final score than LifeSeeker, inspite of having lower precision and recall evaluation metrics.

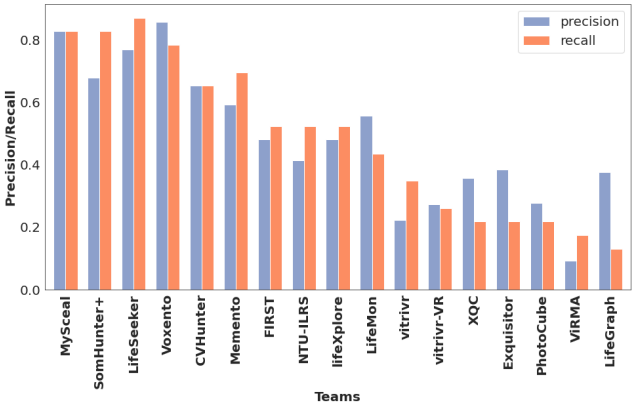


FIGURE 3. Precision and Recall per team across all tasks.

C. SEARCH TIME

The score for each solved task was given based on not only the number of incorrect items previously submitted by that team during that task but also the time taken to find the correct image. We now further investigate the performance of each team by analysing the submission time of systems. Figure 4 illustrates the search time of teams only when they successfully submitted correct answers. This means we exclude the data from tasks that a system could not solve. Because SomHunter+ had the lowest median search time for solved tasks, this system secured the second place although the third-placed LifeSeeker managed to obtain better precision and recall. Nevertheless, MySCéal is the team having a stable and low overall search time. Most of the time when MySCéal found an answer for a task (16 out of 19 solved tasks), it only took them half of the maximum allowed time meaning that they spent less than 150 seconds while 300 seconds was allowed. Contrary to MySCéal, liveXplore and LifeMon are examples of teams that usually submitted their correct answer later than other teams, although the latter system could solved a task with less than a minute twice.

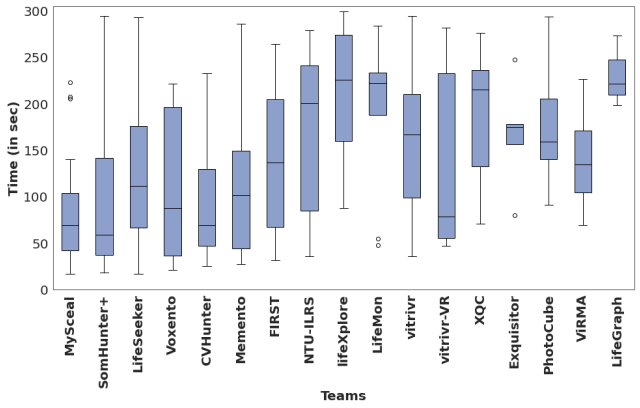


FIGURE 4. Elapsed time until the first correct submission per team across all tasks.

Figure 5 depicts the search time of each system for all tasks regardless of whether they could find the correct answer or not. If a system cannot submit the correct result within this time, no score is allocated for that task. As can be seen for the chart, there is a correlation between the rank of systems with their search time. Systems having lower search time (faster performance) tend to have a better rank. Although sharing the same back-end search engine, Voxento needed slightly less time to retrieve the results than Memento. Recalling the point made earlier about participants usually performing better in second and subsequent years, in this case, Voxento was participating in the second challenge, with Memento participating for the first tine, so this observation holds true in this case. There are other two teams using the same back-end which were vitrivr and vitrivr-VR. Despite solving less tasks (as shown Table 3), vitrivr-VR showed the promise of applying virtual reality environments to lifelog retrieval when they got the correct answer faster than the desktop version for some solved tasks as indicated in Figure 4. Moreover, Figure 5 reveals a large difference in the search time between the top-6 teams and other teams. This is one of the reasons leading to a big gap in the final scores between Memento and FIRST.

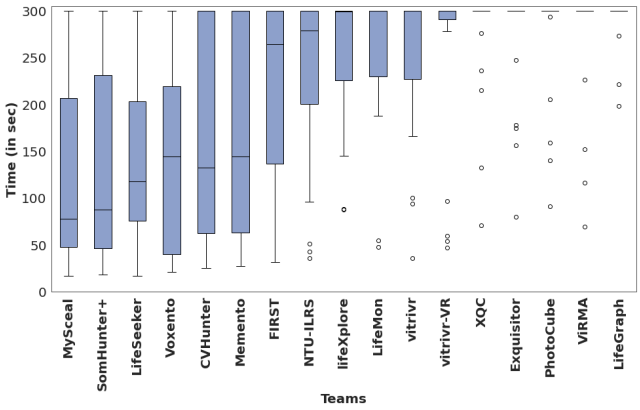


FIGURE 5. Time till first correct submission per team across all task, including unsolved tasks. The 300 second time duration should be taken as an unsolved task.

D. QUERY ANALYSIS

We now analyze more details about each task in LSC'21. Table 4 shows examples of tasks used in the challenge. A task included 6 hints (separated by the character ' / ' as depicted in the Table 4) in which a new hint was shown to the participants every 30 seconds, in order to provide additional contextual information for the searcher. The tasks were constructed with a pattern of initially showing a vague information about the specific images (i.e. '*white t-shirt*') then incrementally revealing more detail (e.g. '*it said I love bicycle*', '*afternoon*'). The final piece of information shown to teams usually indicated the particular places or times of the answers of the tasks (i.e. '*15th May 2015*').

Figure 6 illustrates the score awarded to systems for each query they solved. The maximum score that a team can obtain for each task is 100. The tasks are ordered in the ascending level of difficulty (how hard it was to find the correct answer). The task is considered as easy if many teams can solve it and have high average score. For example, Figure 6 shows that most of teams found the correct answer for task 5 and with relative high score, whilst there were only 4 systems could not solve the task. The reason behind this is that most of the important detail were given in the beginning hints including visual information with places and times ('*whiteboard*', '*student*', '*blue and black top*', '*in the office*', '*in 2016*').

Task 14 was the task that teams managed to solve in the shortest time period. There were 4 teams that almost instantly solved this query just right after the first hint was revealed which were MyScéal, SomHunter+, LifeSeeker, and CVHunter. This is because the very first hint mentioned the '*TagHeuer*' text in the images. There was only a few images containing this brand name, hence this OCR hint helped teams to find the result quickly and nearly get the absolute score for the task. For teams that implemented OCR or the CLIP model (with the inherent OCR), this was an easy task.

One of the hardest tasks was task 11. Only 3 teams which were MyScéal, LifeSeeker, and lifeXplore submitted the correct answer. However, there was a huge difference in their scores where MyScéal achieve 88 but the latter two only scored 54 and 53 accordingly. For this task, even though its hints showed many visual details in the beginning, it was not easy to solve as the main object mentioned in the query text '*telescope*' was not in the provided metadata of the lifelog dataset. Moreover, other things such as '*red flower vase*', '*white violin*', and '*painting*' were not clearly visible in the image that contained the telescope. For very complex queries such as this, an effective interactive retrieval system (which supports browsing of the temporal context of search result) can assist the user in locating the desired items.

Task 4 and Task 18 were also difficult to solve, in that only a few teams managed to get the correct submission. Moreover, the scores they were awarded also considerably lower than other tasks. Regarding task 4, although it had the OCR information in the hints ('*I love bicycle*'), the correct image included a bicycle icon and not the text. Additionally, the images were taken from the distance meaning that the text

TABLE 4. Some example tasks in LSC'21.

Task ID	Hints
4	There was a white t-shirt for sale./ I remember it said I love bicycle./ It was in a bicycle and parts store./ A big sale, bicycles were half price./ It was in the afternoon before I drove to another store then drove to my home/ on the 15th May 2015.
5	Planning a thesis(dissertation) on a whiteboard with my PhD student,/ who was wearing a blue and black stripey top / in my office in 2016./ We were using blue, black and green pens./ After this I went back to work at my computer./ It was on the 27th September.
6	I was roasting marshmallows/ on a BBQ/ at home/ before watching football on the TV./ After roasting the marshmallows I cooked some hamburgers on the same BBQ./ It was in the evening time in May 2018.
11	Telescope in the mirror before going to the airport./ I was able to see my telescope and a red flower vase in the mirror in my bedroom./ I also remember a white violin,/ flowers and a nice painting./ I was playing with my computer and phone at the time./ It was in 2015
14	I remember the TagHeuer advertisement for a watch./ It was a footballer and a watch./ The footballer was sideways kicking the ball./ It was right before I went down stairs in the airport/ in Frankfurt/ in 2015.
18	I was looking at small computer chips on rolls./ It was in a small university electronics laboratory in China./ There were at least 100 rolls of small computer chips./ It was part of a tour of computing and engineering facilities/ and I was with a small delegation of people./ It was in May 2018.

was difficult to read. The remaining hints of this task did not give any additional visual detail to find the answer but until the final clue about the date. In contrast, task 18 consisted of many informative hints. Nevertheless, the '*small computer chips on rolls*' proved hard for searchers conceptualize to search.

It is clear that the queries pose different levels of challenge for teams, and that some systems are better suited to certain query types than other teams. For example, systems that implemented OCR (either explicitly or implicitly by using CLIP embeddings) were better suited to almost half of the queries, while systems that implemented some form of mapping or location-based search would have benefited from the numerous queries that mentioned locations. An additional complexity is that the tasks were created by the lifelogger who gathered the data, and the lifelogger did not consider the eventual distribution of query types.

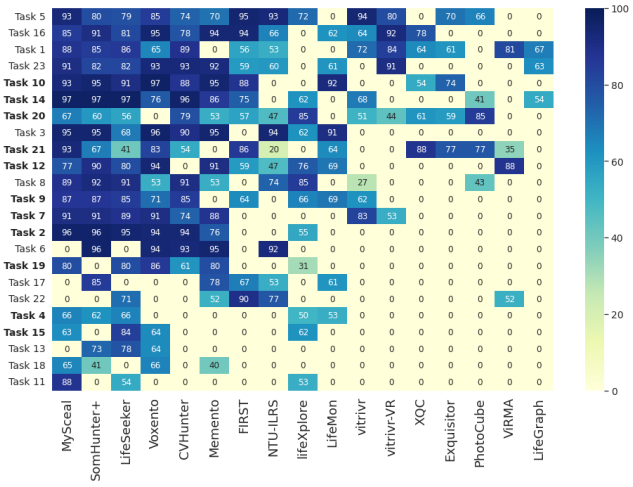


FIGURE 6. Scores of all team across all tasks. The tasks are ordered in respect to the total of scores earned by all teams. The ones written in bold contain OCR hints.

VI. KEY LEARNINGS

The LSC'21 challenge took place at the end of 2021. After four instances of the LSC challenge, and with a focus on the LSC'21 (4th) challenge, we can identify the key features that we consider to be the most promising components of a lifelog retrieval system.

Multimodal embeddings have been introduced in LSC'21 and have shown great potential for bridging the semantic gap between the indexed media content and the textual information needs of the searcher. The free-text search model CLIP has been successfully applied in the image-text retrieval problem [12] now is employed in the lifelog retrieval field by a number of systems, such as SomHunter+, CVHunter, Voxento, and Memento. This embedding approach showed its effectiveness by helping these teams to get high scores, especially the second place of SomHunter+. Furthermore, this model did not require searchers to type only concepts indexed in the search engine, which many searchers would have had no knowledge about. It is anticipated that more groups will employ embedding models for future lifelog search challenges.

Given the temporal nature of lifelog data, it is natural that some queries included temporal cues. Some groups have supported temporal query handling either explicitly in the interface, or by employing basic NLP techniques.

Regarding OCR, half of the tasks in LSC'21 could have benefited from OCR text as a data source. Hence by integrating this feature the systems could solve some queries fast and efficiently, which was rewarded by the LSC scoring algorithm. It should be noted that OCR functionality can be achieved explicitly by using OCR toolkits or implicitly by implementing an embedding model such as CLIP.

The other applied techniques outlined in section IV all have been shown to contribute to the performance of a modern interactive lifelog retrieval system. Some of the techniques, such as location visualisation and novel interaction approaches are all lifelog specific techniques that are implemented by many teams. Others, such as OCR and relevance feedback are more conventional techniques that would be obvious additions to a visual interactive retrieval challenge.

In order to develop a competitive lifelog search system in 2022, it is suggested that teams consider the above mentioned features in their system design as a core set of features, in addition to the conventional aspects of an interactive retrieval system. It is also worth noting that while multimodal embedding models such as CLIP provide a clear benefit to teams, that they will not work for all types of queries and that that additional features will be useful to cover a wide range of query types.

VII. CONCLUSION AND FUTURE PLANS

We have described the 4th annual Lifelog Search Challenge in which 17 teams participated. We provided a short overview of the main features of the systems developed by teams. Although each system was implemented with different approaches for the search engines, all lifelog retrieval systems shared the same goal, which was to find the correct answer in the short manner of time with the least wrong submission. Our analysis showed that LSC'21 witnessed the competitiveness between the top performing teams (MyScéal, SomHunter+, and LifeSeeker) where there was just a small difference in their final scores. We found that the main points differentiated the top teams was not the number of tasks solved (they were equivalent), but the number number of wrong submissions with the fast retrieval time were the critical aspects. This is directly related to the scoring mechanism implemented, which penalised slow or incorrect submissions, however this scoring measure (also implemented in the VBS challenge) represents a good attempt to incorporate many important factors for interactive retrieval into one measure. Therefore a larger-scale evaluation of the top systems with a similar methodology as for video retrieval systems [62] could be interesting.

LSC will continue with larger (multi-year) datasets in future years and with updates to the challenge configuration. The coming LSC challenges will include some new types of tasks [63] such as question answering or ad-hoc queries which requires participants to find all relevant lifelog images rather than only one as the LSC'21. In the following years, the challenge will also explore opportunities for synthetic datasets.

REFERENCES

[1] J. Meyer, S. Simske, K. A. Siek, C. G. Gurrin, and H. Hermens, "Beyond quantified self: Data for wellbeing," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst.*, Apr. 2014, pp. 95–98.

[2] C. Gurrin, K. Schoeffmann, H. Joho, A. Leibetseder, L. Zhou, A. Duane, D.-T. Dang-Nguyen, M. Riegler, L. Piras, M.-T. Tran, J. Lokoc, and W. Hurst, "[Invited papers] comparing approaches to interactive lifelog search at the lifelog search challenge (LSC2018)," *ITE Trans. Media Technol. Appl.*, vol. 7, no. 2, pp. 46–59, 2019.

[3] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal, "NTCIR lifelog: The first test collection for lifelog research," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2016, pp. 705–708.

[4] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, R. Gupta, R. Albatal, D. Nguyen, and D. Tien, "Overview of NTCIR-13 lifelog-2 task," in *Proc. NTCIR*, 2017, pp. 6–11.

[5] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, V.-T. Ninh, T.-K. Le, R. Albatal, D.-T. Dang-Nguyen, and G. Healy, "Overview of the NTCIR-14 Lifelog-3 task," in *Proc. 14th NTCIR Conf.*, 2019, pp. 14–26.

[6] L. Rossetto, R. Gasser, L. Sauter, A. Bernstein, and H. Schuldt, "A system for interactive multimedia retrieval evaluations," in *Proc. Int. Conf. Multimedia Model.*, vol. 12573, J. Lokoc, T. Skopal, K. Schoeffmann, V. Mezaris, X. Li, S. Vrochidis, and I. Patras, Eds. Prague, Czech Republic: Springer, 2021, pp. 385–390.

[7] S. Heller, V. Gsteiger, W. Bailer, C. Gurrin, B. Jónsson, J. Lokoc, A. Leibetseder, F. Mejzlík, L. Peska, L. Rossetto, K. Schall, K. Schoeffmann, H. Schuldt, F. Spiess, L.-D. Tran, L. Vadicamo, P. Veselý, S. Vrochidis, and J. Wu, "Interactive video retrieval evaluation at a distance: Comparing sixteen interactive video search systems in a remote setting at the 10th video browser showdown," *Int. J. Multimedia Inf. Retr.*, vol. 11, no. 1, pp. 1–18, Mar. 2022.

[8] J. Lokoc, F. Mejzlik, P. Vesely, and T. Soucek, "Enhanced SOMHunter for known-item search in lifelog data," in *Proc. 4th Annu. Lifelog Search Challenge*, New York, NY, USA, 2021, pp. 71–73.

[9] L.-D. Tran, M.-D. Nguyen, N. T. Binh, H. Lee, and C. Gurrin, "Myscéal 2.0: A revised experimental interactive lifelog retrieval system for LSC'21," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 11–16.

[10] L.-D. Tran, M.-D. Nguyen, N. T. Binh, H. Lee, and C. Gurrin, "Myscéal: An experimental interactive lifelog retrieval system for LSC'20," in *Proc. 3rd Annu. Workshop Lifelog Search Challenge*, Jun. 2020, pp. 23–28.

[11] J. Lokoc, F. Mejzlik, P. Veselý, and T. Soucek, "Enhanced SOMHunter for known-item search in lifelog data," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 71–73.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.

[13] X. Li, C. Xu, G. Yang, Z. Chen, and J. Dong, "W2VV++: Fully deep learning for ad-hoc video search," in *Proc. 27th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2019, pp. 1786–1794.

[14] T.-N. Nguyen, T.-K. Le, T. Ninh, M.-T. Tran, B. Nguyen, G. Healy, A. Caputo, and C. Gurrin, "Lifeseeker 3.0: An interactive lifelog search engine for LSC'21," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 41–46.

[15] T.-K. Le, V.-T. Ninh, D.-T. Dang-Nguyen, M.-T. Tran, L. Zhou, P. Redondo, S. Smyth, and C. Gurrin, "LifeSeeker: Interactive lifelog search engine at LSC 2019," in *Proc. ACM Workshop Lifelog Search Challenge*, New York, NY, USA, Jun. 2019, pp. 37–40.

[16] A. Alateeq, M. Roantree, and C. Gurrin, "Voxento 2.0: A prototype voice-controlled interactive search engine for lifelogs," in *Proc. 4th Annu. Lifelog Search Challenge*, New York, NY, USA, 2021, pp. 65–70.

[17] A. Alateeq, M. Roantree, and C. Gurrin, "Voxento: A prototype voice-controlled interactive search engine for lifelogs," in *Proc. 3rd Annu. Workshop Lifelog Search Challenge*, Dublin, Ireland, Jun. 2020, pp. 77–81.

[18] N. Alam, Y. Graham, and C. Gurrin, "Memento: A prototype lifelog search engine for LSC'21," in *Proc. 4th Annu. Lifelog Search Challenge*, New York, NY, USA, Aug. 2021, pp. 53–58.

[19] J. Lokoc, F. Mejzlik, T. Soucek, P. Dokoupil, and L. Peska, "Video search with context-aware ranker and relevance feedback," in *MultiMedia Modeling*, B. Jónsson, C. Gurrin, T. Minh-Triet, D. T. Dang-Nguyen, A. M. C. Hu, B. H. T. Thanh, and B. Huet, Eds. Cham, Switzerland: Springer, 2022, pp. 505–510.

[20] M.-T. Tran, T.-A. Nguyen, Q.-C. Tran, M.-K. Tran, K. Nguyen, V.-T. Ninh, T.-K. Le, H.-P. Trang-Trung, H.-A. Le, H.-D. Nguyen, T.-L. Do, V.-K. Vo-Ho, and C. Gurrin, "FIRST–Flexible interactive retrieval SysTem for visual lifelog exploration at LSC 2020," in *Proc. 3rd Annu. Workshop Lifelog Search Challen*, 2020, pp. 67–72.

[21] W.-H. Ang, A.-Z. Yen, T.-T. Chu, H.-H. Huang, and H.-H. Chen, "LifeConcept: An interactive approach for multimodal lifelog retrieval through concept recommendation," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 47–51.

[22] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and caption regions," 2017, *arXiv:1707.09700*.

[23] A. Leibetseder and K. Schoeffmann, "LifeXplore at the lifelog search challenge 2021," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 23–28.

[24] A. C. Faisst and B. Jónsson, "LifeMon: A MongoDB-based lifelog retrieval prototype," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 75–80.

[25] S. Heller, R. Gasser, M. Parian-Scherb, S. Popovic, L. Rossetto, L. Sauter, F. Spiess, and H. Schuldt, "Interactive multimodal lifelog retrieval with vitrivr at LSC 2021," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 35–39.

[26] L. Rossetto, I. Giangreco, C. Tanase, and H. Schuldt, "Vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections," in *Proc. 24th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2016, pp. 1183–1186.

[27] L. Rossetto, R. Gasser, S. Heller, M. Amiri Parian, and H. Schuldt, "Retrieval of structured and unstructured data with vitrivr," in *Proc. ACM Workshop Lifelog Search Challenge*, Jun. 2019, pp. 27–31.

[28] S. Heller, M. Amiri Parian, R. Gasser, L. Sauter, and H. Schuldt, "Interactive lifelog retrieval with vitrivr," in *Proc. 3rd Annu. Workshop Lifelog Search Challenge*, Jun. 2020, pp. 1–6.

[29] R. Gasser, L. Rossetto, S. Heller, and H. Schuldt, "Cottontail DB: An open source database system for multimedia retrieval and analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4465–4468.

[30] R. Gasser, L. Rossetto, S. Heller, and H. Schuldt, "Multimedia retrieval and analysis with cottontail DB," *ACM SIGMultimedia Records*, vol. 13, no. 1, p. 1, Mar. 2021.

[31] L. Rossetto, I. Giangreco, S. Heller, C. Tanase, and H. Schuldt, "Searching in video collections using sketches and sample images—The Cineast system," in *Proc. Int. Conf. Multimedia Modeling*, 2016, pp. 336–341.

[32] F. Spiess, R. Gasser, S. Heller, L. Rossetto, L. Sauter, M. van Zanten, and H. Schuldt, "Exploring intuitive lifelog retrieval and interaction modes in virtual reality with vitrivr-VR," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 17–22.

[33] F. Spiess, R. Gasser, S. Heller, M. Parian-Scherb, L. Rossetto, L. Sauter, and H. Schuldt, "Multi-modal video retrieval in virtual reality with vitrivr-VR," in *Proc. Int. Conf. Multimedia Modeling* in Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2022, pp. 499–504.

[34] F. Spiess, R. Gasser, S. Heller, L. Rossetto, L. Sauter, and H. Schuldt, "Competitive interactive video retrieval in virtual reality with vitrivr-VR," in *Proc. Int. Conf. Multimedia Modeling*, Cham, Switzerland: Springer, 2021, pp. 441–447.

[35] O. S. Khan, A. Duane, B. Jónsson, J. Zahálka, S. Rudinac, and M. Worring, "Exquisitor at the lifelog search challenge 2021: Relationships between semantic classifiers," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 3–6.

[36] O. S. Khan, B. Jónsson, S. Rudinac, J. Zahálka, H. Ragnarsdóttir, H. Þorleiksdóttir, G. Guðmundsson, L. Amsaleg, and M. Worring, "Interactive learning for multimedia at large," in *Proc. Eur. Conf. IR Research (ECIR)*, Lisbon, Portugal: Springer, 2020, pp. 495–510.

[37] E. Knudsen, T. H. Qvortrup, O. S. Khan, and B. Jónsson, "XQC at the lifelog search challenge 2021: Interactive learning on a mobile device," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 89–93.

[38] J. Shin, A. Waldau, A. Duane, and B. P. Jónsson, "PhotoCube at the lifelog search challenge 2021," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 59–63.

[39] B. P. Jónsson, G. Tómasson, H. Sigurþórsson, Á. Eiríksdóttir, L. Amsaleg, and M. K. Lárusdóttir, "A multi-dimensional data model for personal photo browsing," in *Proc. Int. Conf. MultiMedia Modeling (MMM)*, Sydney, NSW, Australia: Springer, 2015, pp. 345–356.

[40] A. Duane and B. P. Jónsson, "ViRMA: Virtual reality multimedia analytics at LSC 2021," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 29–34.

[41] L. Rossetto, M. Baumgartner, R. Gasser, L. Heitz, R. Wang, and A. Bernstein, "Exploring graph-querying approaches in LifeGraph," in *Proc. 4th Annu. Lifelog Search Challenge*, Aug. 2021, pp. 7–10.

[42] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds. Lake Tahoe, NV, USA, 2013, pp. 2787–2795.

[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–14.

[44] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[45] S. Heller, R. Gasser, C. Illi, M. Pasquinelli, L. Sauter, F. Spiess, and H. Schuldt, "Towards explainable interactive multi-modal video retrieval with vitrivr," in *Proc. Int. Conf. Multimedia Modeling*, Cham, Switzerland: Springer, 2021, pp. 435–440.

[46] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[47] J. Devlin, M.-w. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[48] T.-T. Chu, C.-C. Chang, A.-Z. Yen, H.-H. Huang, and H.-H. Chen, "Multimodal retrieval through relations between subjects and objects in lifelog images," in *Proc. 3rd Annu. Workshop Lifelog Search Challenge*, Jun. 2020, pp. 51–55.

[49] S. Heller, L. Rossetto, L. Sauter, and H. Schuldt, "Vitrivr at the lifelog search challenge 2022," in *Proc. 5th Annu. Lifelog Search Challenge*, New York, NY, USA, Jun. 2022, pp. 27–31.

[50] S. Heller, L. Sauter, H. Schuldt, and L. Rossetto, "Multi-stage queries and temporal scoring in vitrivr," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Jul. 2020, pp. 1–5.

[51] S. Heller, R. Arnold, R. Gasser, V. Gsteiger, M. Parian-Scherb, L. Rossetto, L. Sauter, F. Spiess, and H. Schuldt, "Multi-modal interactive video retrieval with temporal queries," in *Proc. Int. Conf. Multimedia Modeling*, 2022, pp. 493–498.

[52] J. Lokoc, G. Kovalcik, T. Soucek, J. Moravec, and P. Cech, "A framework for effective known-item search in video," in *Proc. 27th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2019, pp. 1777–1785.

[53] T.-K. Le, V.-T. Ninh, M.-T. Tran, T.-A. Nguyen, H.-D. Nguyen, L. Zhou, G. Healy, and C. Gurrin, "LifeSeeker 2.0: Interactive lifelog search engine at LSC 2020," in *Proc. 3rd Annu. Workshop Lifelog Search Challenge*, New York, NY, USA, Jun. 2020, pp. 57–62.

[54] M. Kratochvil, F. Mejzlík, P. Veselý, T. Souček, and J. Lokoč, "SOMHunter: Lightweight video search system with SOM-guided relevance feedback," in *Proc. 28th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2020, pp. 4481–4484.

[55] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 20–37, Jan. 2000.

[56] S. Heller, M. Parian, M. Pasquinelli, and H. Schuldt, "Vitrivr-explore: Guided multimedia collection exploration for ad-hoc video search," in *Proc. 13th Int. Conf. Similarity Search Appl. (SISAP)*, Copenhagen, Denmark, vol. 12440, 2020, pp. 379–386.

[57] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[59] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

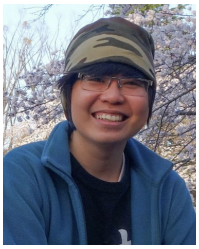
[60] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Jun. 2004.

[61] H.-L. Luo, H. Wei, and L. L. Lai, "Creating efficient visual codebook ensembles for object categorization," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 41, no. 2, pp. 238–253, Mar. 2011.

[62] L. Rossetto, R. Gasser, S. Heller, M. Parian-Scherb, L. Sauter, F. Spiess, H. Schuldt, L. Peska, T. Soucek, M. Kratochvil, F. Mejzlik, P. Vesely, and J. Lokoc, "On the user-centric comparative remote evaluation of interactive video search systems," *IEEE MultimediaMag.*, vol. 28, no. 4, pp. 18–28, Oct. 2021.

[63] J. Lokoč, W. Bailer, K. U. Barthel, C. Gurrin, S. Heller, B. . Jónsson, L. P. ska, L. Rossetto, K. Schoeffmann, L. Vadicamo, S. Vrochidis, and J. Wu, "A task category space for user-centric comparative multimedia search evaluations," in *Proc. Int. Conf. Multimedia Modeling*, Cham, Switzerland: Springer, 2022, pp. 193–204.

LY-DUYEN TRAN received the B.Sc. degree in computer science from the VNU Ho Chi Minh University of Science, in 2019. She is currently pursuing the Ph.D. degree with Dublin City University, doing research into multimodal retrieval from lifelogs. She is also a principal developer of the MySCéal system, which has performed well in all recent instances of the LSC challenge. She is the author of 11 publications and has participated in multiple benchmarking activities with LSC, NTCIR, and ImageCLEF.



MANH-DUY NGUYEN received the B.Sc. degree in computer science from the VNU Ho Chi Minh University of Science, in 2017. He is currently pursuing the Ph.D. degree with the School of Computing, Dublin City University. He is also one of the developers of myscéal system. His research interests include image-text retrieval and graph neural networks.

DUC-TIEN DANG-NGUYEN (Member, IEEE) is currently an Associate Professor in computer science with the Department of Information Science and Media Studies, University of Bergen. His main area of expertise is in multimedia forensics, lifelogging, multimedia retrieval, and computer vision. He is the author or coauthor of more than 100 peer-reviewed and widely cited research papers. He is also a PC member in a number of conferences in the fields of lifelogging, multimedia forensics, and pattern recognition, and a co-organizer of over 40 special sessions, workshops, and research challenges from ACM MM, ACM ICMR, NTCIR, ImageClef, and MediaEval, during the last ten years. He is also the General Chair of MMM 2023.

SILVAN HELLER (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from the University of Basel, where he is currently pursuing the Ph.D. degree with the Department of Mathematics and Computer Science. He is also one of the core developers of Vitrivr. His research interests include multimedia retrieval and analysis systems.

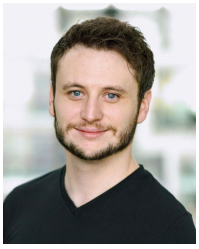
FLORIAN SPIESS received the B.Sc. and M.Sc. degrees in computer science. He is currently pursuing the Ph.D. degree with the Department of Mathematics and Computer Science, University of Basel. His research interests include multimedia retrieval and analysis and applications of virtual reality.

JAKUB LOKOČ is currently an Associate Professor with the Department of Software Engineering, Charles University, Prague. His research interests include similarity search, metric indexing, content-based multimedia analysis, interactive video retrieval, and evaluation of interactive search systems. Recently, he was the General Chair of the MMM 2021 Conference and co-organizes the video browser showdown evaluation campaign.

LADISLAV PEŠKA received the Ph.D. degree from Charles University, Prague, Czech Republic, in 2016. He is currently an Assistant Professor with the Department of Software Engineering, Charles University. He is a member of the SIRET Research Group, Charles University. He also contributes to the development of the CVHunter system. His research interests include recommender systems and multimedia retrieval.

THAO-NHU NGUYEN is currently pursuing the Ph.D. degree in computer science with the School of Computing, Dublin City University. She is also one of the developers of the LifeSeeker system. Her research interests include multimedia analysis, representation, and retrieval, especially for lifelog data.

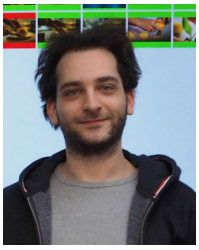
OMAR SHAHBAZ KHAN is currently pursuing the Ph.D. degree with the Department of Computer Science, IT University of Copenhagen. He is also the main developer behind the exquisitor system. His research interest includes scalable multimedia analytics, primarily on large-scale interactive learning for multimedia.



AARON DUANE is currently a Postdoctoral Researcher with the Computer Science Department, IT University of Copenhagen, Copenhagen, Denmark, traveling there from Dublin, Ireland, after being awarded a Marie Curie Postdoctoral Fellowship, in 2020. His research interests include human–computer interaction, multimedia analysis, and virtual reality. For more information visit the link (https://www.linkedin.com/in/aaronduane/).

BJÖRN ÞÓR JÓNSSON is currently an Associate Professor with the Department of Computer Science, Reykjavík University. His research interests include scalable multimedia analytics, scalable multimedia retrieval, and query processing performance in general, especially over novel architectures.

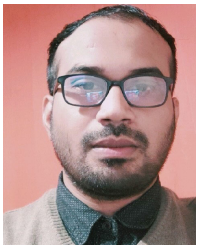
He has recently served as the General Chair for MMM 2022, ACM ICMR 2020, SISAP 2020, and CBMI 2019 conferences, and the Reproducibility Chair for ACM Multimedia (2019–2020) and ICMR (2021–2022). For more information visit the link (http://staff.ru.is/bjorn/).



LUCA ROSSETTO received the Ph.D. degree in computer science from the University of Basel, in 2018. He is currently a Postdoctoral Researcher with the Department of Informatics, University of Zurich. He is also a Core Contributor to the Vitrivr Project and the DRES retrieval evaluation system. His research interests include the analysis, management, and retrieval of multimedia data.

AN-ZI YEN is currently an Assistant Professor with the Department of Computer Science, National Yang Ming Chiao Tung University. Her work has been published in AAAI, SIGIR, WWW, CIKM, and COLING. Her research interests include natural language processing and information retrieval. For more information visit the link (http://nlg.csie.ntu.edu.tw/ azyen/).

AHMED ALATEEQ received the B.Sc. and M.Sc. degrees in computer science from Imam Mohammad Ibn Saud Islamic University, Saudi Arabia. He is currently pursuing the Ph.D. degree with the School of Computing, Dublin City University. He is also a core developer of the voxento system. His research interests include interactive multimedia retrieval systems and multimedia analysis and exploration, especially for lifelog data.



NAUSHAD ALAM received the B.S. and M.S. degrees in computer science from Aligarh Muslim University, India. He is currently pursuing the Ph.D. degree with the School of Computing, Dublin City University. His work has been published at lifelog information retrieval benchmarking challenges hosted at venues, such as ACM ICMR and NTCIR-16. His research interests include multimodal data analytics, lifelogging, information retrieval, and natural language processing.

MINH-TRIET TRAN (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Science, VNU-HCM, in 2001, 2005, and 2009, respectively. In 2001, he joined the University of Science. He was a Visiting Scholar with the National Institutes of Informatics (NII), Japan, from 2008 to 2010, and the University of Illinois at Urbana–Champaign (UIUC), from 2015 to 2016. He is currently the Vice President of the University of Science, VNU-HCM. His research interests include cryptography, security, computer vision, and human–computer interaction. He is a Membership Development and Student Activities Coordinator of the IEEE Vietnam. He is also a member of the Advisory Council for Artificial Intelligence Development of Ho Chi Minh City and the Vice Chairperson of the Vietnam Information Security Association (VNISA, South Branch).

GRAHAM HEALY received the B.Sc. degree (Hons.) in computer applications, in 2008, and the Ph.D. degree in brain–computer interfaces, in 2012. He is currently an Assistant Professor with the School of Computing, Dublin City University. His research interests include data analytics and human–computer interaction, with a particular focus on novel interaction paradigms and data indexing/processing strategies.

KLAUS SCHOEFFMANN received the M.Sc. and Ph.D. degrees in computer science. He is currently an Associate Professor with the Institute of Information Technology (ITEC), Klagenfurt University, Austria. He is also one of the developers of lifeXplore. His research interests include video content understanding (in particular medical/surgery videos), multimedia retrieval, interactive multimedia, and applied deep learning.

CATHAL GURRIN received the B.Sc. and Ph.D. degrees in computer applications from Dublin City University, in 1997 and 2002, respectively.

He is currently the Co-Founder of the LSC Lifelog Search Challenge Workshop and the NTCIR-Lifelog collaborative benchmarking activity. He is the coauthor of *LifeLogging: Personal Big Data*. His research has been covered internationally on the BBC, Discovery Channel, and in print media (The Economist and The New York Times).