Paper

Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC2018)

Cathal Gurrin [†1], Klaus Schoeffmann [†2], Hideo Joho [†3],
Andreas Leibetseder [†2], Liting Zhou [†1], Aaron Duane [†1],
Duc-Tien Dang-Nguyen [†4], Michael Riegler [†5], Luca Piras [†6],
Minh-Triet Tran [†7], Jakub Lokoč [†8], Wolfgang Hürst [†9]

Abstract   The Lifelog Search Challenge (LSC) is an international content retrieval competition that evaluates search for personal lifelog data. At the LSC, content-based search is performed over a multi-modal dataset, continuously recorded by a lifelogger over 27 days, consisting of multimedia content, biometric data, human activity data, and information activities data. In this work, we report on the first LSC that took place in Yokohama, Japan in 2018 as a special workshop at ACM International Conference on Multimedia Retrieval 2018 (ICMR 2018). We describe the general idea of this challenge, summarise the participating search systems as well as the evaluation procedure, and analyse the search performance of the teams in various aspects. We try to identify reasons why some systems performed better than others and provide an outlook as well as open issues for upcoming iterations of the challenge.

Key words: lifelog, collaborative benchmarking, interactive retrieval, evaluation

1. Introduction

Technological progress over the last decade and the ready availability of low-cost sensors means that individuals can now capture detailed traces of their life experience, which are commonly referred to as lifelogs. Initially, driven by a desire for self-knowledge to enhance personal health and wellness[1], a range of novel life-experience sensors, such as wearable cameras, or audio recording devices, can now passively generate continuous archives of multimodal life experience data in a process called lifelogging. In this work, we assume a definition of lifelogging as introduced by Dodge

and Kitchen[2] which refers to the gathering of 'a unified digital record of the totality of an individual's experiences, captured multimodally through digital sensors and stored permanently as a personal multimedia archive'. Such sensors can include wearable camera or audio sensors to capture everyday activities from the point-of-view of the wearer, biometric sensors for physical markers of the body, activity sensors for human movement, contextual sensors (e.g. GPS) for context logging, informational sensors (e.g. software) to capture information accesses, and potentially many others. These multimodal datasets pose new challenges for our existing approaches to multimedia information organisation and retrieval[3].

It is our belief that the current generation of multimodal information retrieval systems are not designed to operate effectively with such lifelog archives, which are deeply multimodal, continuous and potentially error-laden[3]. In the spirit of Memex[4], it is our conjecture that a lifelog, if it is to be useful to the individual, must be '*continuously extended, it must be stored, and above all it must be consulted*'. Such lifelog consultation is likely to require both ad-hoc and interactive retrieval mechanisms to support a wide variety of lifelog use-cases, as outlined in both[5] and [3]. While we note significant efforts being made through various vehicles,

such as NTCIR[6] and ImageCLEF[7], to support off-line ad-hoc search tasks, until the Lifelog Search Challenge (LSC*) in 2018, there was no dedicated benchmarking effort for interactive lifelog search. We know from previous efforts for conventional text and multimedia retrieval that such open collaborative benchmarking efforts contribute significantly to advances in domain knowledge[8].

In this work, we highlight advances in the state-of-the-art for interactive lifelog retrieval by collating and reviewing the six interactive retrieval systems developed for the first collaborative benchmarking exercise for lifelog information retrieval (LSC 2018), which took place at the ACM ICMR 2018 conference in Yokohama, Japan in June 2018. The main contribution of this paper is therefore, an comparative review of the performance of six different interactive lifelog retrieval systems on the only dataset ever designed for interactive lifelog retrieval[9] and introducing a novel interactive benchmarking experiment and comparative scoring model.

2. Related Research Activities

The field of information retrieval has a long history of benchmarking exercises in which numerous systems and techniques to solve specific retrieval challenges are compared against each other by using the same test collections openly and cooperatively. Typically this works by participants developing systems, evaluating them over test collections and then (after-the-fact) coming together for an open comparison of system performance. This is best exemplified by the test collection methodology employed by large-scale international efforts, such as TREC[10], CLEF[11], NTCIR[12] and in the multimedia field, efforts such as ImageCLEF[13] or MediaEval[14]. A summary of these activities and their challenges can been found at[15].

2.1 Interactive Benchmarking Exercises
However, most of these efforts do not focus on benchmarking interactive retrieval systems. One related effort that does, however, is the the Video Browser Showdown (VBS[16]), which is an annual international video search competition with the goal to evaluate the state-of-the-art performance of interactive video retrieval systems on a large shared dataset of video data. It has been held as a special session at the International Conference on Multimedia Modeling (MMM), annually since 2012. In this competition several teams work in front of a shared screen and try to solve a given set of retrieval tasks as fast as possible. The tasks are issued and scored by the VBS server, which evaluates the search time and correctness of each submission and computes a score for the team. The whole competition consists of expert and novice sessions, where for the latter, volunteers from the conference audience work with the tools of the experts. The final score is computed as an average over all sessions.

While lifelog retrieval is different from video retrieval, which is the focus of the VBS, both topics have a lot of similarities. Both lifelog archives, and digital video archives are forms of multimodal data archive with temporally organised large datasets (more details can be found in[17]). Whereas video archives typically contain curated and non-errorsome data in two modalities, lifelog datasets are genuinely multimodal by nature, with the strong potential for errors, missing or misaligned data. Consequently, the LSC Challenge, discussed in this paper, is modeled on the successful VBS, though with different aims, dataset and information needs.

2.2 Interactive Lifelog Retrieval Systems
While there are numerous data organisation and retrieval systems designed for lifelog data, in this discussion we focus on interactive systems (i.e. more than query/submit pairs) for multimodal lifelog data archives. The seminal MyLifeBits[18] project at Microsoft produced, what is generally regarded as the first interactive lifelog retrieval system, which was based on a database indexing and retrieval metaphor. Lee et al.[19] went beyond the database metaphor by developing an interactive event-organised lifelog browsing interface for visual lifelog data that segmented days into events, based on analysis of visual and sensor data, and linked events together in a single diary-style interface. More recently, the LEMoRe[20] system, an interactive lifelog retrieval engine, developed in the context of the Lifelog Semantic Access Task (LSAT) of the the NTCIR-12 challenge, integrated classical image descriptors with high-level semantic concepts and was powered by a graphical user interface that uses natural language processing to process a user's query.

While all of these are good examples of interactive lifelog retrieval systems, until LSC 2018, it was not possible to draw any performance comparisons between

---

them. Each of them operated on different (or proprietary) datasets. The LEMoRe system was the only one to index a reusable and publicly available test collection, though no other interactive retrieval engine was available for comparison at that time. Hence, the importance of the LSC 2018, the first opportunity to benchmark approaches to interactive lifelog retrieval, which attracted seven participating groups, although only six actually competed the evaluation, which are described in this paper.

3. LSC 2018 - The Search Challenge

As stated, the LSC 2018 took place during ACM ICMR 2018, in Yokohama, Japan. The LSC was a public competition during which all attendees at the conference were welcome to attend the event and observe the competition. LSC 2018 employed the LSC dataset, which we will now briefly introduce.

3.1 LSC Dataset

The LSC dataset was a 27-day multimodal lifelog dataset gathered by one individual who wore multiple sensors and utilised smartphone and computer software to capture a continuous 24/7 lifelog. Details of the dataset and a description of the methodology employed in the construction of the dataset is described elsewhere[9]. The lifelog data was temporally aligned to UTC time (Coordinated Universal Time) and in order to maintain privacy of the lifelogger and bystanders in the data, all visual content was filtered firstly by the lifelogger themselves and then by a trusted expert, to remove any potentially embarrassing or problematic data. This data was then enhanced by the addition of various forms of metadata before all user identifiable content (e.g. faces, name badges, addresses) was removed and the collection made available for download.

In summary, the dataset consists of:

- Multimedia Content. Wearable camera images (1024 x 768 resolution) were gathered at a frequency of about two images per minute (from breakfast to sleep). Accompanying the wearable camera images were a set of concept annotations generated by the Microsoft cognitive services (computer vision API)[21]. Additionally, a timestamped record of music listening activities sourced from Last.FM* was also included.

- Biometric Data. Human biometrics, such as heart rate, galvanic skin response, calorie burn and steps,

on a per-minute basis were included in addition to daily blood pressure and blood glucose levels (manually recorded every morning before breakfast) and weekly cholesterol and uric acid levels.

- Human Activity Data. Physical activities on a per-minute basis (e.g. walking, running, standing), a location log of locations visited, along with a time-stamped diet-log of all food consumed drinks taken.

- Information Activities Data. Using the Loggerman app[22], the information creation and consumption activities on a per minute basis, which were organised into blacklist-filtered and alphabetically sorted document vectors representing every minute.

This dataset was represented as a set of JPG images and an XML file with metadata entries for every minute. The data is available for download (after signing-up for access) from the LSC website.

3.2 Topics & Relevance Judgements

In over to facilitate interactive retrieval and competitive benchmarking in a live setting, a novel set of temporally enhanced queries were generated by the lifelogger who gathered the dataset. Each topic was created by the lifelogger selecting a memorable and interesting event that had occurred during the time period covered by the test collection. In total there were six development topics, six test topics for experts (system developers), and twelve test topics for novice users, who were not knowledgeable about the collection or how the systems worked. Only the development topics were released before the competition.

These queries were textual (e.g. 'find when I was in a Norwegian furniture store'), but they were constructed to provide additional contextual information (i.e., get easier) every thirty seconds (e.g. 'I was looking at chairs', 'It was a Monday afternoon'). The topics were temporally extended through six iterations during the live search challenge, with each iteration lasting for 30 seconds and providing increasing levels of contextual data to assist the searcher. With six iterations in total, this resulted in total time allocation of three minutes per topic. Examples of the topics are shown in the Task Presentation section below.

Relevance judgements were generated manually by the lifelogger. There could be one or more relevant items in the collection, where relevant items could span multiple separate events or happenings. In this case, if a user of an interactive system found any one of the relevant items from any event, then the search is deemed

---
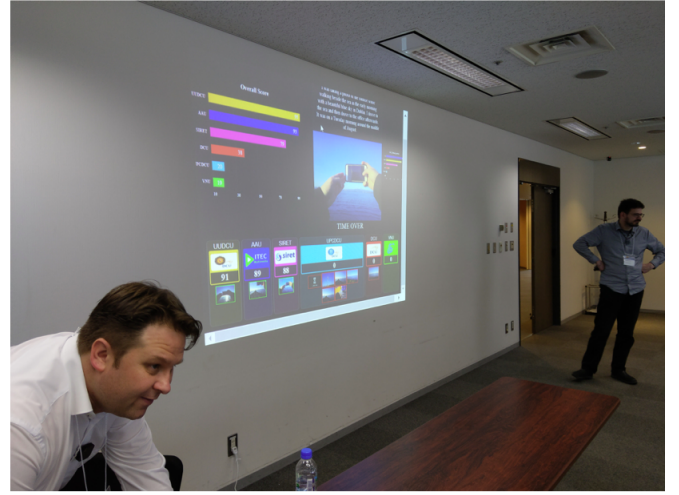
Fig. 1 LSC2018 Competition showing the teams and the live scoreboard

to be successful. For the LSC collection, an item was assumed to be an image from the wearable camera.

### 3.3 Scoring in the Interactive Search Challenge

During the search challenge, participating teams were asked to submit a relevant item to a host server when a potentially relevant item from the collection was found by the participant. The host server maintained a count-down clock and actively evaluated submissions against the groundtruth. Throughout the competition, an over-all score was maintained for each team, which was the summation of the scores of the topics that had been processed up until that point. For each topic, a score was given based on the time taken to find the relevant content and the number of incorrect items previously submitted by that team to the host server during that topic. Full details of the scoring equation are given in the section 'Evaluation of System Performance' below.

4. Participating Teams

In 2018, six participating teams took part in the live search challenge. These teams had all indexed the dataset prior to attending the workshop and then dur-ing the interactive search challenge, both expert and novice users took part in evaluating the performance of the six systems. For the challenge, each participant was given a desk with a clear view of a large screen which showed the topics, the time remaining on each topic, as well as the current and overall scores of each team. The physical configuration of the challenge can be seen in Figure 1.

We explore the results in more detail in a later sec-tion, but firstly we highlight the six approaches taken by the participating teams.
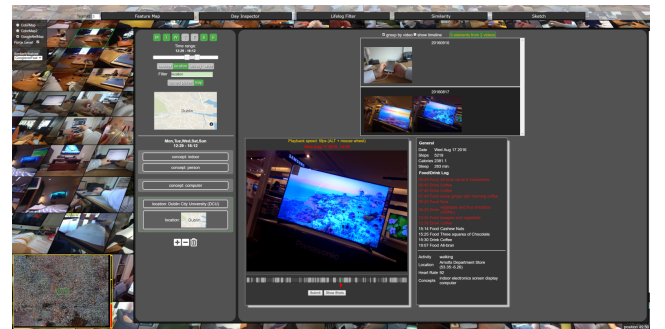

Fig. 2 liveXplore Interactive Interface from AAU

4.1 AAU: liveXplore at the Lifelog Search Challenge 2018

The successful employment of the web technologies-based diveXplore system[23] by Alpen-Adria-Universität Klagenfurt (AAU) at past iterations of the annual Video Browser Showdown led to the development of liveXplore[24], a system modification serving as a lifel-ogging data browser by focusing on visual exploration and retrieval as well as metadata filtering. Since the ap-plication is developed for processing video scenes, LSC image sequences were converted to video using a con-stant frame rate. Pre-calculated semantic shot segmen-tation enabled clustering of similar images to coherent scenes and the creation of the main interface, which presented the user with an adjustable multi-level fea-ture map grouping together similar shots according to machine learning descriptors or handcrafted features. Additionally to providing shot-specific similarity search based on these features, liveXplore specifically offered
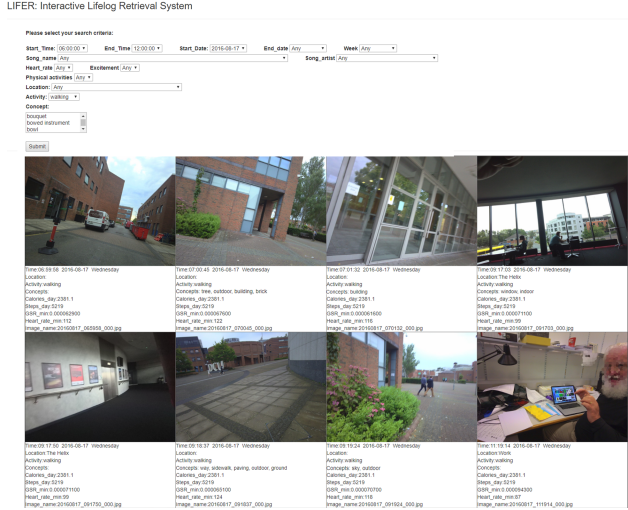
Fig. 3　The LIFER Retrieval Interface from DCU

Fig. 4　The UPC-DCU Interactive Retrieval Interface

the possibility of exploring individual lifelog day summaries as chronologically ordered galleries as well as videos in an overlay view enriched with metadata information. Finally, in order to search the data according to metadata information the system featured a filter view that allowed users to mix and match temporal, location- or activity-based and machine learning concept oriented filtering. The liveXplore interface is shown in Figure 2.

While filtering options such as the selection of day-time, weekday, activity, named location and provided machine learning concepts proved to be very useful for finding correct scenes, others were identified as less useful: heart rate, skin temperature as well as exact geolocation. This, of course, can be attributed to the current rather small dataset magnitude and variety, thus, potentially making these options relevant for future LSC iterations, likely to exhibit more data from several different sources. Future liveXplore versions will comprise further promising filtering options, specifically focusing on non-metadata related exploration.

4.2　DCU: LIFER, An Interactive Lifelog Retrieval System

Dublin City University (DCU) took part with a first generation interactive lifelog search engine called LIFER[25], a system that allows a user to retrieve the moments from the personal life archives in a fast and efficient manner. The LIFER system was designed to assist a user in examining their life experience to gain insights into their activities and lifestyle. LIFER was developed to index only the locations, concepts, time, and activities from the provided dataset, which were the features that the developers felt would provide most benefits in an interactive setting. This data was converted
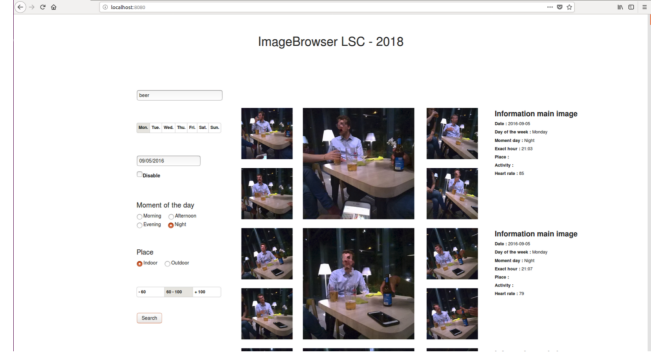
into feature vectors over every minute. These feature vectors were hierarchically grouped into event nodes. The retrieval is then performed by collected moments (in this task, images) that matched with the queried criteria and presenting them on screen in a ranked list with associated metadata, as shown in Figure 3. Selecting any image allows it to be submitted to the server for judgement.

Queries were submitted as sets of facets relating to date / time, biometrics, activities, locations, visual concepts and music consumed. These facets were merged to generate feature vectors for similarity ranking.

4.3　UPC-DCU: Interactive Lifelog Image Browser

The Interactive Lifelog Browser developed by Universitat Politecnica de Catalunya (UPC) in collaboration with Dublin City University (DCU), was a novel retrieval engine based on three core considerations: (1) the development of a multi-faceted query interface, (2) the inclusion of a trusted retrieval engine, and (3) the novel presentation of a ranked list of results[26].

Borrowing from the standard WWW-interface for faceted search systems (e.g. hotel booking or fight booking), the interface was designed with two sections, as shown in Figure 4. On the left side the query panel is displayed which contains the faceted and free-text query elements. On the right side is the result display panel. The faceted search components included Day-of-the-Week selector, Calendar selector, moment-of-the-day selector (time of day), Place selector and Heart-rate.

The ranking engine indexed every minute as the retrievable unit using the commonly used TF-IDF ranking methodology. The free text search implements standard enhancements, such as stopword removal and term stemming for the English language. This ranked list from the free-text search is filtered by the other data
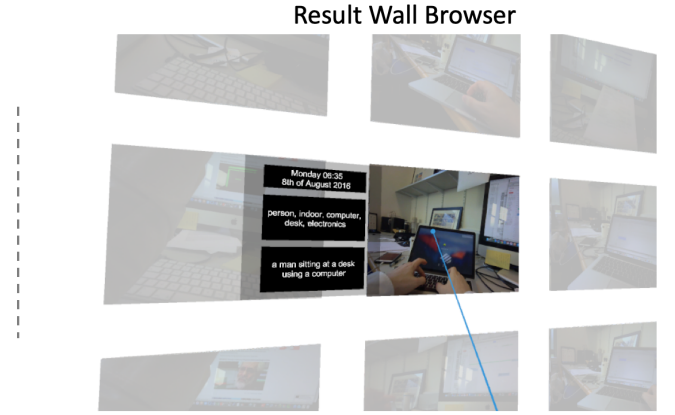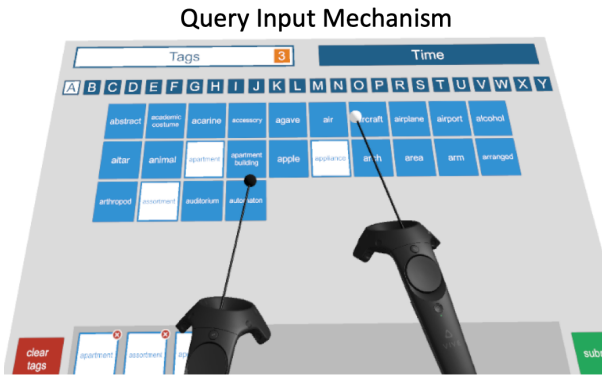
Fig. 5 UU-DCU Virtual Reality Interface showing the query and browsing mechanisms

facets, such as time of day, day of week, or location. The result is a ranked list of filtered moments for presentation to the user. In order to provide the user with some context of a ranked moment, the previous two images and the following two images contribute (on a sliding scale) to the overall score of the main image. Selecting an image allows it to be submitted to the server for judgement.

### 4.4 UU-DCU: Virtual Reality Lifelog Explorer

The virtual reality lifelog explorer developed for the LSC in a collaboration between University of Utrecht (UU) and Dublin City University (DCU) has two components, each of which needed to be optimised for a VR environment[27]. The querying component was a virtual interface designed to provide a quick and efficient means for a user to generate a filter query within the VR system. This gesture-based querying interface consisted of two sub-menus, one for selecting lifelog concepts of interest and the second for selecting the temporal aspect of the query (e.g. hours of the day or days of the week). Only these two sources of evidence were used in the VR Explorer. A contact-based approach was employed, which utilised a direct form of interaction where the user must physically touch the interface elements with their controllers, which required a drumstick-like appendage protruding from the head of each controller in the VR environment (see Figure 5, left-side). Tactile feedback was provided through the hand-controllers to signify hitting the buttons.

After a filter query is submitted to the system, the querying interface disappeared, and the user was presented with the highest-ranked filtered images in decreasing rank order, in a left-to-right organised result wall. The ranking was based on a combination of concept relevance and the time of capture (maintaining the temporal organisation of the data), where concept relevance took precedence over the temporal arrangement. Any image displayed on the VR ranked list could be selected for further exploration by pointing the user's controller at it and pressing a button (see Figure 5, right-side). This showed additional metadata about the image such as the specific capture date and time and what concepts have been detected. Other filtering options were also made available along with this metadata. For example, the user had the option of viewing all the images captured before and after the target image within a specific timespan. Upon finding a potentially relevant image, the user could submit it to the LSC server for validation and scoring.

4.5 VNU-HCM: Semantic Concepts Fusion Retrieval

The group from the University of Science and University of Information Technology (Vietnam National University-Ho Chi Minh city) developed a pioneering lifelog retrieval system that integrated recent achievements in computer vision for place and scene attribute analysis, object detection and localization, and activity detection using image captioning[28][29]. This system can be highlighted according to the three main novel advancements: (1) Visual Clustering for Images: independent images are organised into visual shots, sequences of similar images, based on visual information, then visually similar sequences are linked to a scene using visual retrieval with Bag-of-Word framework, (2) Concept Extraction: the system extracts the location of as well as the scene attributes of an image and create a textual caption of the image for indexing, (3) Augmented Data
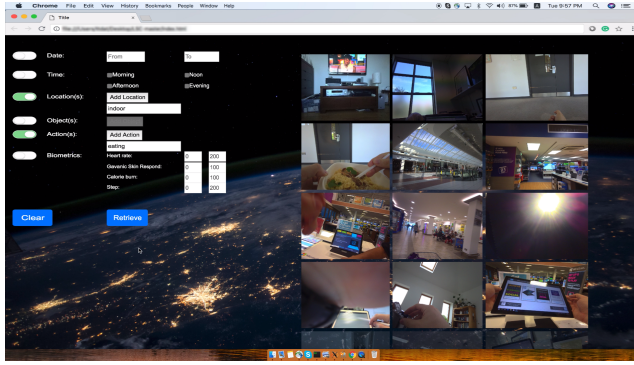
Fig. 6 VNU-HCM Interactive Retrieval System

Fig. 7 The VIRET Lifelog Retrieval Tool from SIRET

Processing: besides visual information, lifelogging data also contain useful augmented data, such as biometrics, blood pressure, blood sugar level, text data of computer activities, etc. Indices were created for such augmented data in an indexing process.

The system provided four groups of search features corresponding to four different groups of query criteria: (1) Temporal criteria: a user can specify the date and time, time span, or period (morning, afternoon, etc), (2) Scene criteria: a user can specify a query on scene categories (hotel, restaurant, lobby, etc) or scene attributes (open area, camping, sunbathing, etc), (3) Entity and Action criteria: a user can specify a query on the existence of entities, or actions/activities, (4) Extra criteria: a user can define a query on biometrics data, computer usage information, etc.

The overall interface allowed the user to integrate all of these core techniques in one comprehensive system, as shown in Figure 6 with the query panel on the left and the result panel on the right.

4.6 SIRET: VIRET - An Interactive Lifelog Search Engine

After a successful participation at the Video Browser Showdown 2018 (1st place), the SIRET team from Charles University, Prague, participated also at the Lifelog search challenge with an updated version of the VIRET system[30]. The objective of the participation was to inspect the performance of a purely content-based video retrieval tool for Lifelog data. The tool did not consider provided lifelog specific modalities (e.g., locations or heart rate). Since the tool relies on sequences of extracted video frames, the transition to the visual Lifelog repository was straightforward. Every day from the collection was treated as one 'video' represented by the lifelog images, extended by selected images/frames extracted from provided short videos. For each image,

automatic annotations were obtained from a retrained GoogleNet (with an own set of 1,390 ImageNet labels). In addition, a colour signature for sketch-based search and deep feature vector from the original GoogleNet were extracted. Based on the automatically extracted features, users could provide three types of query input (keywords, colour sketch and example images) that could be further combined by a late fusion strategy. More specifically, each modality could be used to define a subset of top relevant images and the intersection of all constructed subsets was returned as the result. The final result list was sorted by selected modalities and displayed in the presentation panel. The VIRET tool supported two types of result presentation – classical grid with images sorted by relevance and a result list enhanced with nearby temporal context for each top matching frame. Whereas the grid with more images is useful for exploration phase of the search with frequent query reformulation actions, the temporal context view helps with inspection of promising (visually similar) candidates. To inspect a temporal context in the grid, users can display all images from the corresponding day in the bottom panel. In addition, the mouse wheel can be used to quickly inspect the temporal context of each displayed image (the images change in the grid cell). Even though the tool performed relatively well (the overall third place), it turned out that the additional Lifelog modalities would be important for effective filtering. Therefore, we plan to incorporate the modalities in the future versions of the VIRET tool. The VIRET interface is shown in Figure 7.

4.7 Comparison of System Features

Table 1 shows a basic comparison between features implemented in each system. Some features were expected to provide obvious utility to developers, such as the facet filters which were employed in some form by

Table 1	Summary of the Features used by all six Participating Systems)					
Feature	AAU	DCU	UPC-DCU	UU-DCU	VNUHCM-US	VIRET
Facet Filters	Y	Y	Y	Y	Y	Y
Event/Scene Organisation	Y	Y	N	N	Y	Y
Visual Clustering	Y	N	N	N	Y	N
Novel Ranked List Visualisation	Y	N	Y	Y	N	Y
Enhanced Visual Analytics	Y	N	N	N	Y	Y
Integration of Biometric Data	Y	Y	N	N	Y	N
Non-textual/faceted Querying Mechanism	Y	N	N	N	Y	Y
Based on Existing Video Search Tool	Y	N	N	N	N	Y

all systems. Most systems also incorporated some form of event/scene organisation in the user interface, as well as producing a novel form of ranked list in response to a user information need. Interestingly, only half of the systems actually implemented biometric filters as part of the query process. Finally, we note that two of the systems (liveXplore and VIRET - two of the top three ranked systems) were based on existing video browsing/retrieval systems, which were refined to work with lifelog data.

5. Evaluation of System Performance

To better understand the evaluation procedure of the LSC challenge, we describe how tasks are presented and how novice and expert users differ. The expert users would typically be the system developers themselves, while novice users are recruited from the audience of the conference and are expected not to be familiar with any internal details of the system. We assume that experts would be faster than novices who had not seen the system before the challenge. Integrating novices into the competition is important because it supports the goal of the LSC, which is to foster research into user-friendly lifelog search systems. This goal is also the reason why at LSC 2018 we tested more tasks with novice users than with experts.

5.1 The LSC Server

Similar to the Video Browser Showdown (VBS)[31] the Lifelog Search Challenge uses a dedicated server software on-site (the *LSC Server*) to present task descriptions and evaluate submissions on-the-fly. Whenever a team submitted an answer to the HTTP-based server, it would immediately respond with an indication whether the submission was correct or not. Furthermore, it would also display the evaluation results (correct or wrong; topic scores and overall scores) on a scoreboard, such that other teams and the audience will be notified when some team has found/submitted a segment for verification and be aware of the overall

scores of the teams.

5.2 Calculating Scores

At LSC 2018 we issued 18 temporal queries Q that were separated into 6 *expert* and 12 *novice* tasks ($Q = \{E \cup N\}$). The participants were required to solve these queries as fast and accurately as possible, as they got points for each task dependent on the required search time and the number of wrong submissions.

As shown in Equation 1, for every team t the task score S_q^t of a task q is computed based on the maximum achievable points A_q for that task (we used $A_q = 100$ for every task), the search time τ_q^t required by the team to solve the task, the number of wrong submissions for the task ω_q, and the maximum provided search time T_q for the task (which varied among experts and novices, as described below). This scoring is designed such that the score linearly decreases from the maximum to half of the points over the allowed search time (and will be zero in worst case).

Therefore, if a task will count 100 points and a team is able to find the correct segment in the last second without any wrong submissions, it will still get 50 points. However, for every wrong submission the basis for this linear decrease will lower to 90 percent of the current basis, such that for the same situation but with two wrong submissions, the team will only get 31 points (and with five wrong submissions only 9.05 points). Thus, it is quite important to verify the correctness of the retrieved segment before submitting it to the LSC server for scoring.

$$S_q^t = max(0, A_q \cdot \frac{T_q \cdot 0.9^{\omega_q} - 0.5 \cdot \tau_q^t}{T_q}) \qquad (1)$$
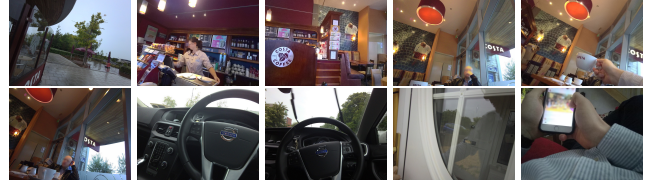
The preliminary team score for the expert session S_E^t and the novice session S_N^t is computed as the sum of all task scores in the session, as given in Equations 2 and 3.
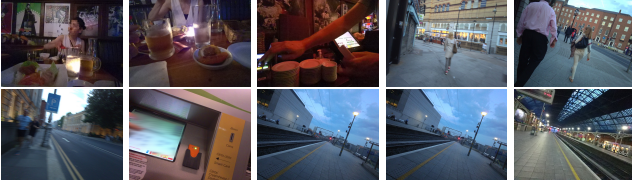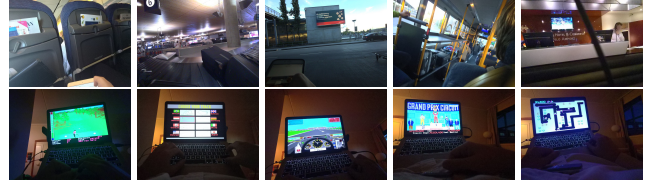
$$S_E^t = \sum_q^E S_q \qquad (2)$$
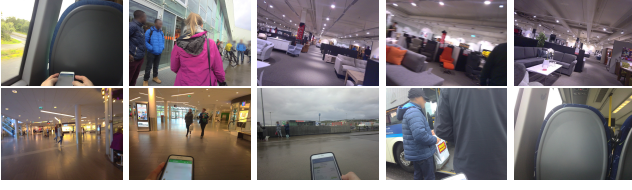
Fig. 8   Images of Expert Task E01

Fig. 9   Images of Expert Task E05

Fig. 10   Images of Novice Task N01

Fig. 11   Images of Novice Task N05

$$S_N^t = \sum_q^N S_q \tag{3}$$

Finally, the maximum team score per session (M_E and M_N) is determined and used to normalise all preliminary team scores of each session to compute the final points P^t for each team:

$$P^t = \frac{S_E^t}{M_E} + \frac{S_N^t}{M_N} \tag{4}$$

This way we end up with an achievable maximum of 200 points as the final result for a team that scored best in both expert and novice sessions.

5.3   Task Presentation

Tasks (textual descriptions) are projected onto a large screen by the LSC Server. Each task is represented by the temporal query, which is textual in nature and incrementally refined after every 30 seconds. For example, the first expert task at LSC 2018 started with "*I was in a Norwegian furniture store in a shopping mall...*". After 30 seconds the query description was extended with "*...where I was looking at chairs.*". After one minute even more details were added ("*There is a large 'SALE' or 'SALG' sign in the store.*") and after 30 more seconds some specific time information was presented: "*It is a Monday afternoon.*". This scheme of incrementally extending the query is repeated exactly five times until the full query was available (i.e., the last extension was provided after two minutes and 30 seconds). This is true for both the expert and the novice tasks.

5.4   Expert Tasks at LSC2018

Overall, at LSC2018 six tasks had to be solved by the experts, who got a time limit of only three minutes (180 seconds). In the following list you can see the final text

of two example expert tasks and the first few images of the ground truth from the life logger (Figure 8-9):

E01   . "*I was in a Norwegian furniture store in a shopping mall where I was looking at chairs. There is a large 'SALE' or 'SALG' sign in the store. It is a Monday afternoon. I went to the store by bus and I took a bus to a restaurant after I finished in the mall.*" For examples, see Figure 8.

E05   "*I was waiting for the train in Dublin city after walking to the station from a sushi restaurant where I had dinner and beer by candlelight. It was on a Tuesday night and I ate in a restaurant called Yamamori.*" For examples, see Figure 9.

5.5   Novice Tasks at LSC2018

For novice users twelve tasks had to be solved, each with a time limit of five minutes (300 seconds). Please note that we used the same number of query refinements, i.e., after 02:30 no more extensions to the query were presented, but the participants had more time to find the relevant content. In the following list you can see two example topics from the novice tasks, including images of the ground truth data (Figure 10-11).

N01   "*There was a large picture of a man carrying a box of tomatoes beside a child on a bicycle. I was having Saturday morning Coffee in Costa Coffee with a friend, the first in September. After coffee I drove home and played with my phone. Coffee began about 8am and finished about 9:35am.*" For examples, see Figure 10.

N05   "*I was playing a vintage car-racing game on my laptop in a hotel after flying to Norway. I played a number of different types of vintage computer game before and after the car-racing game. It was in the evening on a Saturday in a Clarion airport hotel. I took a bus to the hotel from the airport.*" For examples, see Figure 11.
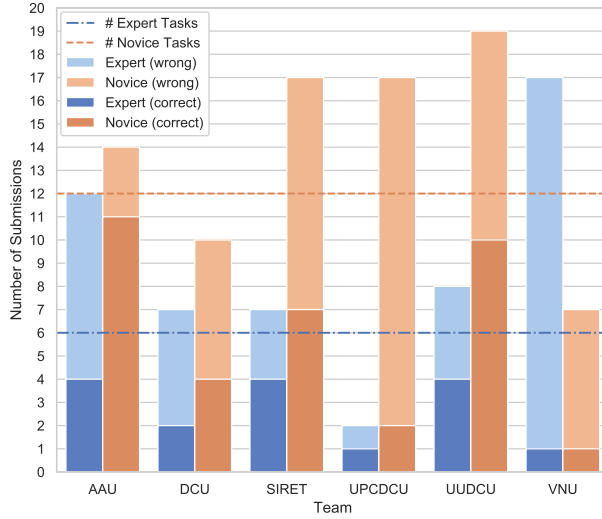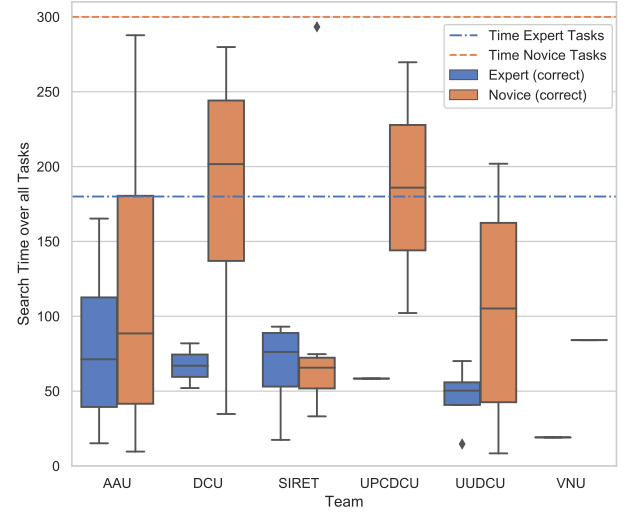
Fig. 12 Number of correct/wrong submissions per team and user across all tasks.

Fig. 13 Distribution of search times for correct submissions over all tasks per team and user type.

5.6 Number of Correct/Wrong Submissions

In order to analyse the performance of the teams, we will inspect their submissions first. Figure 12 shows the number of correct and wrong submissions over all tasks, separated into expert and novice groupings. As can be seen, no team could solve every task in the expert session, but AAU, SIRET, and UU-DCU solved four out of six (actually no team could solve the very first task shown above - E01). Among these three teams SIRET and UU-DCU had a similar number of wrong submissions (3 vs. 4), while AAU submitted about twice as many wrong ones (i.e., 8). UPC-DCU and VNU could only solve one expert task, but VNU submitted a lot of wrong submissions (i.e., 16), which would have reduced their scores significantly.

When looking at the novice session, we can see that AAU and UU-DCU could solve almost all twelve tasks (11 vs. 10), while DCU, UPC-DCU, and VNU could only solve a few (4, 2, and 1). It is also apparent that AAU had significantly less wrong submissions – in relation to the correct submissions – than in the expert session (only 3/14 vs. 8/12), while for UPC-DCU and SIRET this relation was significantly higher (15/17 and 10/17 vs. 1/2 and 3/7). We believe that this was caused by variability in the ability and expertise of the novice users.

In total over both sessions, AAU solved most tasks (15 out of 18) and VNU solved least (only 2 out of 18). However, in order to determine the best team we also need to look at the search time, which is analysed in the next section.

5.7 Search Time

Figure 13 presents a box-plot for the search time (in seconds) over all tasks for all teams. In general we can observe that the novices required more time to find the correct scene and also had a larger variation than experts, but they were also required to solve twice as many tasks. However, this general observation is not true for SIRET, for who we can see a similar search time for experts and novices (actually, the median search time of novices is even lower than the one of experts). This suggests that the SIRET interface is intuitive for both novices and experts alike. Additionally, Figure 14 shows the search time of a correct submission per task and team (over all eighteen tasks). This figure again demonstrates the higher task solving performance of AAU and UU-DCU, who could solve almost all tasks. AAU even solved more than UU-DCU (15 vs. 14), but UU-DCU was much more efficient in terms of search time – which is also the reason why they could finally win the competition. DCU and UPC-DCU could only solve a few tasks and required a relatively long time to find the relevant content. SIRET is somewhat in-between and VNU unfortunately could solve only two tasks, but with a good search time when their system performed well.

5.8 Total Score/Points Calculation

As discussed above, the winner of the LSC competition is determined by normalizing the scores of both sessions to the maximum score of each session. Figure 15 shows the result of this normalisation. UU-DCU achieved the best score in the expert session and got 100 points (followed by SIRET and AAU with 90.56
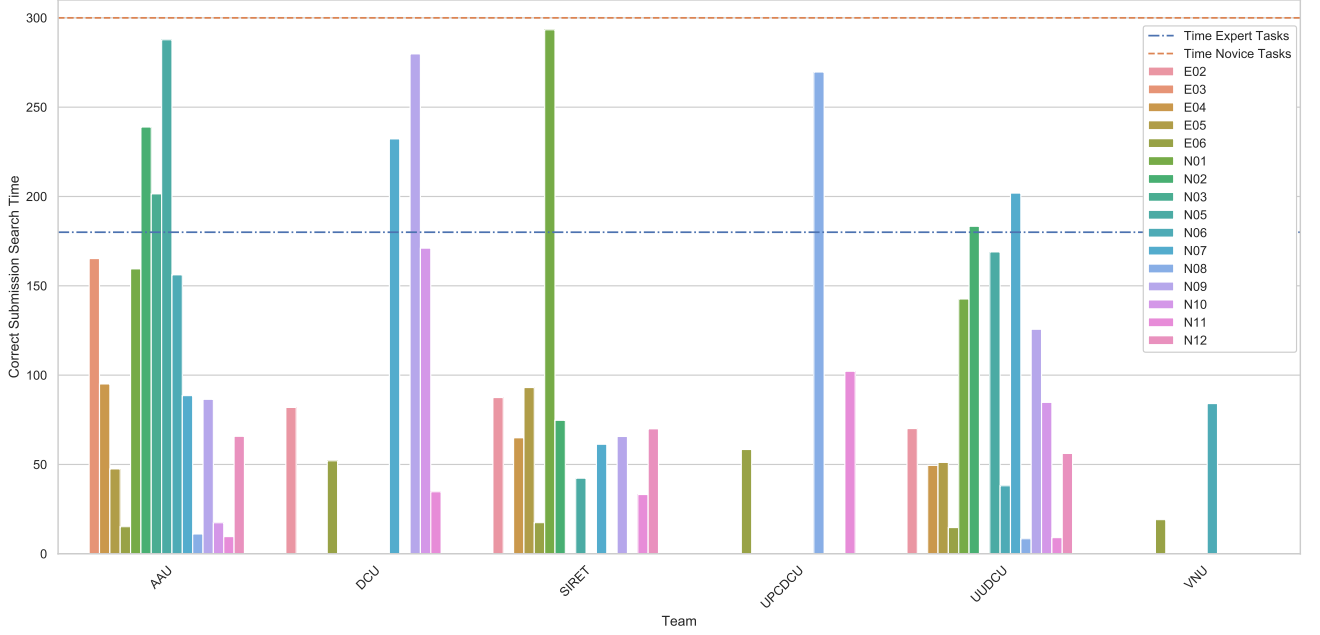
Fig. 14 Correct submission times per team and task (Expert and Novice).

and 85.84 points). In the novice session the situation was similar but with a different winner: AAU got 100 points for the best score and was followed by UU-DCU and SIRET with 89.06 and 67.54 points. Thus, the overall winner of the LSC2018 competition is UU-DCU with a total of 189.06 points. There was a significant gap to the bottom ranked three teams. Later observation suggested that, although these systems used similar indexed data, their performance was hampered by other issues, such as system performance in a competition environment, or errors in the system implementation.
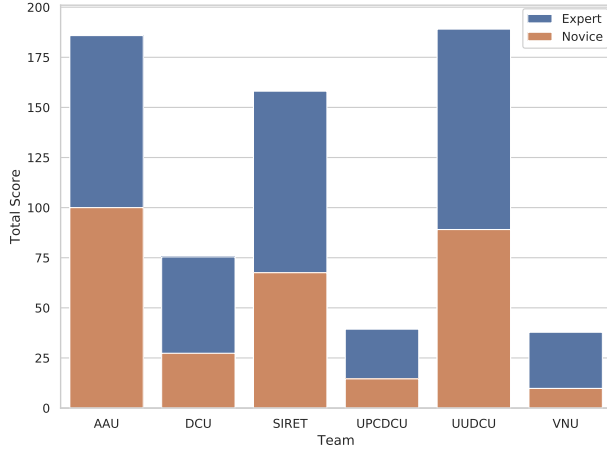
6. Discussion

With only six participants, and given that this is the first time to run the Lifelog Search Challenge, it is difficult to identify clear reasons as to why one system outperforms all others. However we can make some observations. The top three performing systems (UU-DCU, AAU and SIRET) were all able to utilise existing retrieval systems that had been developed to address other tasks and challenges, thereby reducing the potential for technical difficulties. It is no surprise therefore that all three of these systems performed well with no technical problems. AAU and SIRET were based on existing systems that have successfully competed in, and won at the Video Browser Showdown[16] in recent years. UU-DCU which performed marginally better overall than AAU was based on an existing lifelog browsing

system developed over a number of years previously. Examining the results (15) in critical detail, the difference in the scores between UU-DCU and AAU were marginal, though it is notable that AAU performed better in the novice task, which is likely a more fair reflection of actual system performance, when the expert user has been removed from the evaluation.

There is one final point that should be noted, given the short duration of the dataset (27 days) and the fact that the dataset was released to participants many months in advance of the competition, there is always the potential for an expert user, who is familiar with the dataset to gain an advantage over other users. However, it is likely that any potential learning effect would have been the same overall participants; thus it is not understood if this had any impact on system performance.

In terms of the relative performance of experts and novices, the differences in performance between both types of user is clearly illustrated in Figure 16. Expert users typically found relevant content faster than novice users across all topics. Novices took significantly longer than the expert users. However, these observations need to be considered with a little caution, since novices solved twice as many tasks than experts.

It is not clear whether the inclusion of biometric metadata and other activity data sources helps much in the interactive retrieval process. As shown in Table 1, only three of the systems integrated such data into their ranking processes, but there is no clear indication
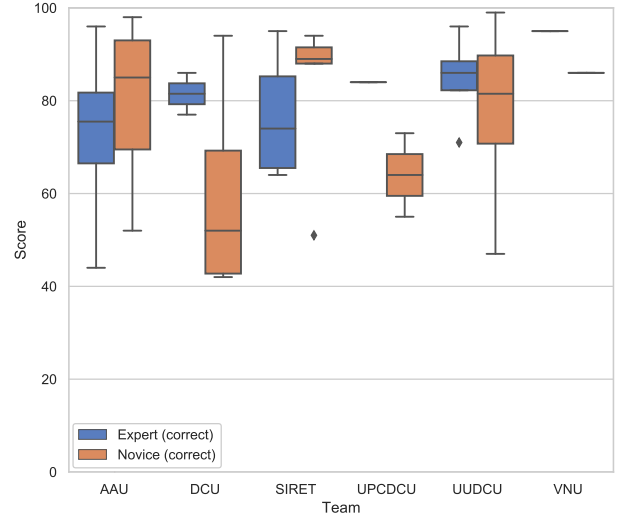
Fig. 15 Total score of teams by user (normalized).

Fig. 16 Distribution of task score for correct submissions over all tasks per team and user.

as to whether this data helped or hindered the process.

Finally, in terms of complexity of system design, the three systems (DCU, UU-DCU, and UPC-DCU) all integrated only the provided dataset and metadata and developed their retrieval systems over this data. SIRET, AAU, and VNU-HCM, on the other hand, have applied some enhancements to the dataset based on their experience with video retrieval. This insight suggests that the techniques shown to be effective in competitions such as the VBS did not transfer readily to the LSC dataset. It appears that indexing lifelog data will require the development of multimodal lifelog-specific toolkits to enhance performance beyond a baseline level which all three top-performing teams have met in the first LSC. Future editions of the LSC will shed more light on such issues and bring the community closer to a consensus on how best to support an individual to interactively locate data from massive multimodal lifelogs, which is a topic that the LSC organisers consider to be an increasingly important research topic as society edges closer to an era in which large-scale personal lifelogs becomes the norm, rather than the exception.

7. Conclusions and Future Plans

In this paper, we presented an overview of the first Lifelog Search Challenge (LSC 2018), that was organised at ACM ICMR 2018, in Yokohama, Japan. Six participating teams took place in the competition, each of which developed and utilised an interactive lifelog search engine. In this first edition of the LSC, we note that there was a clear distinction between the three top performing teams and those that ranked less highly. The best performing teams had re-purposed existing in-

teractive retrieval systems to operate with multimodal lifelog data, two of which had applied additional multimedia analytics tools to extract additional metadata. As to be expected after the first LSC challenge, a good baseline approach for interactive lifelog retrieval is not yet clearly defined, but it appears as if a well tested interactive system, placing significant emphasis on the visual element of lifelog data is a good starting point. The second LSC (2019) will take place at ICMR 2019 in Ottawa, Canada (using the same dataset as LSC 2018) and a third is planned for ICMR 2020 in Dublin, Ireland. The organisers anticipate that clear retrieval strategies will emerge over the coming years as more LSC challenges are run.

8. Acknowledgements

References

1) Jochen Meyer, Steven Simske, Katie A. Siek, Cathal G. Gurrin, and Hermie Hermens. Beyond quantified self: Data for wellbeing. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pages 95–98, New York, NY, USA, 2014. ACM.

2) Martin Dodge and Rob Kitchin. 'Outlines of a world coming into existence': Pervasive computing and the ethics of forgetting. *Environment and Planning B: Planning and Design*, 34(3):431–445, 2007.

3) Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. LifeLogging: Personal Big Data. *Foundations and Trends® in Information Retrieval*, 8(1):1–125, 2014.

4) Vannevar Bush. As we may think. *interactions*, 3(2):35–46,

March 1996.

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

6) Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Rashmi Gupta, Rami Albatal, and Duc-Tien Dang-Nguyen. Overview of ntcir-13 lifelog-2 task. In *Proceedings of NTCIR-13*, 2017.

7) Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. Overview of image-cleflifelog 2018: Daily living understanding and lifelog moment retrieval. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, 2018.

8) Alan F. Smeaton, Paul Over, and Wessel Kraaij. Trecvid: Evaluating the effectiveness of information retrieval tasks on digital video. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pages 652–655, New York, NY, USA, 2004. ACM.

9) Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Bernd Munzer, Rami Albatal, Frank Hopfgartner, Liting Zhou, and Duc-Tien Dang-Nguyen. A Test Collection for Interactive Lifelog Retrieval. In Ioannis Kompatsiaris, Benoit Huet, Vasileios Mezaris, Cathal Gurrin, Wen-Huang Cheng, and Stefanos Vrochidis, editors, *MultiMedia Modeling*, pages 312–324, Cham, 2019. Springer International Publishing.

10) Ellen M. Voorhees and Angela Ellis, editors. *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*, volume Special Publication 500-324. National Institute of Standards and Technology (NIST), 2017.

11) Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier, editors. *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

12) Yiqun Liu, Makoto P. Kato, Charles L. A. Clarke, Noriko Kando, and Tetsuya Sakai. Report on NTCIR-13: the thirteenth round of NII testbeds and community for information access research. *SIGIR Forum*, 52(1):102–110, 2018.

13) Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew Lungren, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. Overview of ImageCLEF 2018: Challenges, Datasets and Evaluation. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 309–334, Cham, 2018. Springer International Publishing.

14) Martha Larson, Piyush Arora, Claire-Hélène Demarty, Michael Riegler, Benjamin Bischke, Emmanuel Dellandréa, Mathias Lux, Alastair Porter, and Gareth J. F. Jones, editors. *Working Notes Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29-31 October 2018*, volume 2283 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

15) Duc-Tien Dang-Nguyen, Michael Riegler, Liting Zhou, and Cathal Gurrin. Challenges and opportunities within personal life archives. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 335–343. ACM, 2018.

16) Jakub Lokoc, Werner Bailer, Klaus Schoeffmann, Bernd Münzer, and George Awad. On influential trends in interactive video retrieval: Video browser showdown 2015-2017. *IEEE Trans. Multimedia*, 20(12):3361–3376, 2018.

17) Duc-Tien Dang-Nguyen, Liting Zhou, Rashmi Gupta, Michael Riegler, and Cathal Gurrin. Building a disclosed lifelog dataset: challenges, principles and processes. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, page 22. ACM, 2017.

18) Jim Gemmell, Gordon Bell, and Roger Lueder. Mylifebits: a personal database for everything. *Commun. ACM*, 49(1):88–95, 2006.

19) Hyowon Lee, Alan F. Smeaton, Noel E. O'Connor, Gareth Jones, Michael Blighe, Daragh Byrne, Aiden Doherty, and Cathal Gurrin. Constructing a SenseCam visual diary as a media process. *Multimedia Systems*, 14(6):341–349, 2008.

20) Gabriel de Oliveira Barra, Alejandro Cartas Ayala, Marc Bolaños, Mariella Dimiccoli, Xavier Giró Nieto, and Petia Radeva. Lemore: A lifelog engine for moments retrieval at the ntcir-lifelog lsat task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, 2016.

21) Alessandro Del Sole. *Microsoft Computer Vision APIs Distilled: Getting Started with Cognitive Services*. Apress, 2017.

22) Zaher Hinbarji, Rami Albatal, Noel O'Connor, and Cathal Gurrin. Loggerman, a comprehensive logging andvisualization tool to capture computer usage. In Qi Tian, Nicu Sebe, Guo-Jun Qi, Benoit Huet, Richang Hong, and Xueliang Liu, editors, *MultiMedia Modeling*, pages 342–347, Cham, 2016. Springer International Publishing.

23) Manfred Jürgen Primus, Bernd Münzer, Andreas Leibetseder, and Klaus Schoeffmann. The itec collaborative video search system at the video browser showdown 2018. In Klaus Schoeffmann, Thanarat H. Chalidabhongse, Chong Wah Ngo, Supavadee Aramvith, Noel E. O'Connor, Yo-Sung Ho, Moncef Gabbouj, and Ahmed Elgammal, editors, *MultiMedia Modeling*, pages 438–443, Cham, 2018. Springer International Publishing.

24) Bernd Münzer, Andreas Leibetseder, Sabrina Kletz, Manfred Jürgen Primus, and Klaus Schoeffmann. lifexplore at the lifelog search challenge 2018. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, LSC '18, pages 3–8, New York, NY, USA, 2018. ACM.

25) Liting Zhou, Zaher Hinbarji, Duc-Tien Dang-Nguyen, and Cathal Gurrin. Lifer: An interactive lifelog retrieval system. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, LSC '18, pages 9–14, New York, NY, USA, 2018. ACM.

26) Adrià Alsina, Xavier Giró, and Cathal Gurrin. An interactive lifelog search engine for lsc2018. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, LSC '18, pages 30–32, New York, NY, USA, 2018. ACM.

27) Aaron Duane, Cathal Gurrin, and Wolfgang Hürst. Virtual reality lifelog explorer: Lifelog search challenge at acm icmr 2018. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, LSC '18, pages 20–23, New York, NY, USA, 2018. ACM.

28) Thanh-Dat Truong, Tung Dinh-Duy, Vinh-Tiep Nguyen, and Minh-Triet Tran. Lifelogging retrieval based on semantic concepts fusion. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, LSC '18, pages 24–29, New York, NY, USA, 2018. ACM.

29) Viet-Khoa Vo-Ho, Quoc-An Luong, Duy-Tam Nguyen, Mai-Khiem Tran, and Minh-Triet Tran. Personal diary generation from wearable cameras with concept augmented image captioning and wide trail strategy. In *Proceedings of the Ninth International Symposium on Information and Communication Technology*, SoICT 2018, pages 367–374, New York, NY, USA, 2018. ACM.

30) Jakub Lokoč, Tomáš Souček, and Gregor Kovalčik. Using an interactive video retrieval tool for lifelog data. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, LSC '18, pages 15–19, New York, NY, USA, 2018. ACM.

31) Jakub Lokoč, Gregor Kovalčík, Bernd Münzer, Klaus Schöffmann, Werner Bailer, Ralph Gasser, Stefanos Vrochidis, Phuong Anh Nguyen, Sitapa Rujikietgumjorn, and Kai Uwe Barthel. Interactive search or sequential browsing? a detailed analysis of the video browser showdown 2018. *ACM Trans. Multimedia Comput. Commun. Appl.*, 15(1):29:1–29:18, February 2019.

Cathal Gurrin is an associate professor at the School of Computing, Dublin City University and a principal co-investigator at the Insight Centre for Data Analytics. His research interests are personal analytics and lifelogging and he is researching the next-generation of digital records for the individual. He is a co-founding organiser of many benchmarking challenges for lifelog data including NTCIR-Lifelog, ImageCLEF Lifelog and the Lifelog Search Challenge and he has co-authored more than 200 papers in the general domain of information access and retrieval.

Klaus Schoeffmann is an associate professor at the Institute of Information Technology (ITEC) at Klagenfurt University, Austria. His research focuses on medical multimedia systems, video understanding, and interactive multimedia. He has co-authored more than 110 publications on various topics in multimedia and he has co-organised several international conferences, workshops, and special sessions in the field of multimedia. Furthermore, he is co-founder of the Video Browser Showdown (VBS) an international live evaluation competition of interactive video search.

Hideo Joho is an Associate Professor at Faculty of Library, Information and Media Science, University of Tsukuba, Japan. His research interests are cognitive and affective analysis of interaction between search applications and their users. Recent work includes conversational search, collaborative search, and lifelog search evaluation. Dr. Joho is a Program Co-Chair of CHIIR 2019, General Co-Chair of SIGIR 2017, and Co-Organiser of NTCIR Lifelog Tasks.
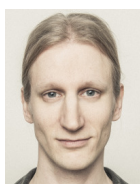
Andreas Leibetseder is a PhD student at the Institute of Information Technology (ITEC) at Klagenfurt University, Austria. He is project assistant in the field of medical multimedia analysis and his research interests include computer vision, machine learning and human-computer interaction.
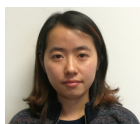
Liting Zhou is a Phd candidate at the Insight Centre for Data Analysis in Dublin City University Ireland. Her research focuses on developing information access and retrieval approaches for lifelog data and she is co-founder of the NTCIR Lifelog and ImageCLEF Lifelog collaborative benchmarking activities.
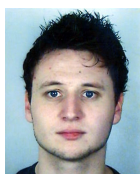
Aaron Duane is a PhD candidate at the Insight Centre for Data Analytics in Dublin City University Ireland. The focus of his research is on visual access to lifelog and personal data in virtual environments.

Duc-Tien Dang-Nguyen is an associate professor at the Department of Information Science and Media Studies at the University of Bergen, Norway. His research interest are multimedia verification, multimedia analytics and personal lifelogging. Within the field of lifelogging, he coordinates the three most important bench-marking initiatives: co-founder of ImageCLEF Lifelog and Lifelog Search Challenge (LSC), and co-organiser of NTCIR-Lifelog. He also holds position on numerous editorial boards, review committees and conference steering committees.

Michael Riegler Michael Riegler is a senior researcher at Simula Research Laboratory. He received his Masters degree from Klagenfurt University with distinction and finished his PhD at the University of Oslo in two and a half years. His research interests are medical multimedia data analysis and understanding, image processing, image retrieval, parallel processing, crowdsourcing, social computing and user intent. He is involved in several initiatives like the MediaEval Benchmarking initiative for Multimedia Evaluation and he is part of an expert group for the Norwegian Council of Technology on Machine Learning for Healthcare.

Luca Piras is a post-doc researcher at the University of Cagliari, his research interests include content-based image retrieval and classification, relevance feedback techniques and unbalanced learning. He has been involved as organizer in the ImageCLEF evaluation campaign since 2015 before in the Image Annotation task and then as co-organizer of the Lifelog task. He is co-founder of Pluribus One a spin-off of the University of Cagliari where is currently responsible for the R&D of multimedia and vision products.

Minh-Triet Tran (is the Deputy Head of Software Engineering Department and the Assistant Dean of Honors Program, Scientific & International Affairs at Ho Chi Minh University of Science, Vietnam. His research interests include Cryptography/Information Security, Multimedia & Augmented Reality and Software Engineering. He has written numerous books and on these topics and he was recognised as one of the ten most outstanding youths in Information and Communication Technologies in Vietnam.)

Jakub Lokoč is an assistant professor with the Department of Software Engineering, Faculty of Mathematics and Physics, Charles University. His research interests include metric indexing, multimedia databases, video retrieval, known-item search, and similarity modeling.

Wolfgang Hürst Wolfgang Hürst is an assistant professor at the Department of Information and Computing Sciences at Utrecht University, The Netherlands. His research interests include mobile computing, human-computer interaction, augmented and virtual reality, and multimedia systems and technologies. His research have been published in various articles in high ranked scientific publications. He is a member of the editorial boards of IEEE MultiMedia and IEEE Transactions on Multimedia, and co-chair of various conferences.