Interactive Multimodal Lifelog Retrieval with vitrivr at LSC 2021

Silvan Heller
University of Basel
Switzerland
silvan.heller@unibas.ch

Ralph Gasser
University of Basel
Switzerland
ralph.gasser@unibas.ch

Mahnaz Parian-Scherb
University of Basel, Switzerland
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
mahnaz.parian-scherb@unibas.ch

Sanja Popovic
University of Basel
Switzerland
sanja.popovic@stud.unibas.ch

Luca Rossetto
University of Zurich
Switzerland
rossetto@ifi.uzh.ch

Loris Sauter
University of Basel
Switzerland
loris.sauter@unibas.ch

Florian Spiess
University of Basel
Switzerland
florian.spiess@unibas.ch

Heiko Schuldt
University of Basel
Switzerland
heiko.schuldt@unibas.ch

ABSTRACT

The Lifelog Search Challenge (LSC) is an annual benchmarking competition for interactive multimedia retrieval systems, where participating systems compete in finding events based on textual descriptions containing hints about structured, semi-structured, and/or unstructured data. In this paper, we present the multimedia retrieval system vitrivr, a long-time participant to LSC, with a focus on new functionality. Specifically, we introduce the image stabilisation module which is added prior to the feature extraction to reduce the image degradation caused by lifelogger movements, and discuss how geodata is used during query formulation, query execution, and result presentation.

CCS CONCEPTS

• **Information systems → Search interfaces**; **Image search**; **Users and interactive retrieval**; *Multimedia databases*; **Information retrieval**; • **Human-centered computing → Interactive systems and tools**.

KEYWORDS

Content-based Retrieval, Multimedia Retrieval, Lifelogging, Lifelog Search Challenge

ACM Reference Format:

Silvan Heller, Ralph Gasser, Mahnaz Parian-Scherb, Sanja Popovic, Luca Rossetto, Loris Sauter, Florian Spiess, and Heiko Schuldt. 2021. Interactive Multimodal Lifelog Retrieval with vitrivr at LSC 2021. In *Proceedings of the 4th Annual Lifelog Search Challenge (LSC '21), August 21, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3463948.3469062

LSC '21, August 21, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8533-6/21/08…$15.00
https://doi.org/10.1145/3463948.3469062

1 INTRODUCTION

Ubiquitous mobile computing devices equipped with increasingly powerful and diverse sensors endow users with rich opportunities to digitally capture their individual experiences. While this is a boon for the lifelogging community, enabling the recording of increasingly detailed lifelogs, it also comes with the added difficulty when searching for relevant information in this growing amount of data. Such retrieval capabilities are especially important when the lifelogs are supposed to serve as a memory aid.

The diverse and at the same time personal nature of the data contained within a lifelog, however, makes it difficult to quantify the usefulness of any retrieval approach in an automated fashion. Still, the most effective way for such retrieval evaluations comes in the form of interactive campaigns, which comparatively evaluate retrieval system using common settings and standardized infrastructure [21] on realistic retrieval tasks. Modelled after the Video Browser Showdown (VBS) [16, 20, 27], the Lifelog Search Challenge (LSC) [6–8, 18] is such a campaign, and has established itself as a driving force behind the advances in lifelog retrieval.

For this fourth edition of the LSC, we present in this paper the improvements made to the vitrivr multimedia retrieval stack [23] in the context of its third participation in this series of events. In particular, we review existing functionality for lifelog retrieval, and introduce new functionality included for LSC 2021. Specifically, we describe image stabilization, which we use to improve spatial feature extraction and potentially image browsing. Furthermore, we elaborate on how we use geodata in the entire querying pipeline.

Since vitrivr is a general-purpose system not specifically built for the LSC, some changes and additions to vitrivr during the last year are not discussed in this paper. In summary, we have made changes to Cottontail DB [4, 5], experimented with alternative frontends [11] and used the new competition server DRES [21] to evaluate vitrivr in-depth [19].

The remainder of this paper is structured as follows: Section 2 gives an overview of vitrivr's system architecture and describes its functionality. Section 3 describes the additions to vitrivr for LSC 2021 and in Section 4 we discuss the dataset preparation. Section 5

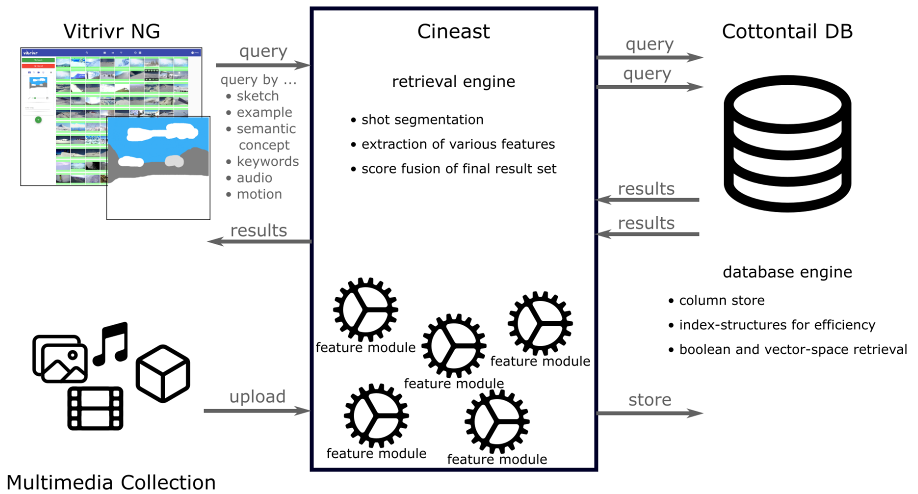

Figure 1: System overview for vitrivr and its three major components: vitrivr-ng, Cineast and Cottontail DB [10].

gives a comprehensive overview of other systems at LSC 2020 and Section 6 concludes.

2 VITRIVR

vitrivr is a general-purpose open-source multimedia retrieval system, supporting various multimedia formats (e.g., video, images, audio, and 3D models) and query modalities (e.g., query-by-sketch, query-by-concept). vitrivr has been a long-time participant at both LSC [9, 18] and VBS [10, 24, 25] with a relatively stable architecture consisting of three components: a database for multimedia features, Cottontail DB, the feature generation and retrieval engine Cineast, and a web-based front-end, vitrivr-ng. In this section, we will give a brief overview of the individual components and their functionality for lifelog retrieval.

2.1 Architecture

Figure 1 shows a system overview of the vitrivr stack. It consists of three major components, which reflect the clear separation of concern we try to implement in the stack:

Database: Cottontail DB is the database management system for multimedia retrieval offering support for both Boolean and similarity search [4]. In the context of LSC, we mainly rely on Cottontail DB's ability to find exact matches for metadata lookup (e.g., range queries) as well as full-text search.

Retrieval Engine: Cineast Cineast is the feature extraction and query processing engine of the vitrivr stack [22]. Its modular architecture enables it to capture and combine various notions of similarity, as represented by different feature representations, both within and across multiple media modalities.

Frontend: vitrivr-ng Written in TypeScript using the Angular[1] framework, vitrivr-ng is responsible for query formulation, result presentation, browsing and filtering. Users can combine various query modalities, for example, first filtering by time and then making a color sketch. Figure 2 shows a screenshot of the current UI.

2.2 Existing Functionality for Lifelog Retrieval

vitrivr has already participated to the LSC in previous years [9, 18] and has thus added functionality specifically for the lifelogging setting. In this section, we briefly summarize the extensions made for previous LSC participations.

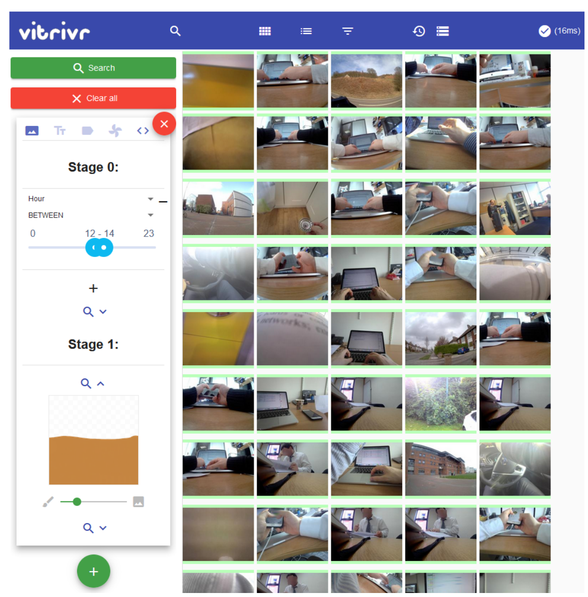

Figure 2: Illustrative screenshot of the user interface with a staged query.

[1] https://angular.io

Image Sequence Media Type. Cineast's data model is centered around *media objects* that can further be sub-divided into *segments*. Features and metadata can then either belong to a media object or a segment. Originally, Cineast only supported simple media types as media objects, such as single images, videos or audio files. Since LSC is centered around sequences of images that belong to the same day, we have added a new media type called an *image sequence* to accommodate these requirements. Similarly to how videos consist of different shots, image sequences contain multiple images that belong together, i.e., the sequence itself is the media object while the individual images are regarded as segments. Features can then be generated on a segment level (e.g., location of an individual image) or on an object level (e.g., day of the week).

Boolean Queries. The original version of Cineast mainly offered support for similarity search and fulltext search on certain units of information, such as extract scene text or speech. Even though it was always possible to store metadata on both a media object and a segment level, that data was usually not queried.

For LSC 2019, we added explicit support for Boolean queries to Cineast, supporting classical operators such as greater / less than, equals and between. These can be used to search in the provided metadata such as location information, weekday, year, etc. Based on that, later versions have also added support for combining Boolean and similarity search using *staged queries* [12] (see fig. 2).

Result Filters. In contrast to content-based multimedia retrieval, which is often based on the notion of similarity, LSC provides us with binary predicates such as *"It was a Saturday in 2016"*. We make use of this in two ways: First, staged queries allow us to apply such filters already during the query formulation phase, allowing the retrieval engine to seamlessly combine Boolean predicates with similarity search . Second, to avoid re-execution of the query once new information appears, we have also added late filtering based on all the metadata provided along with query results, which exclusively takes place in the front-end. This feature could be used to quickly reduce the result set as new information becomes available during the competition, without the need for another round-trip to the retrieval engine.

3 NEW FUNCTIONALITY FOR LSC 2021

3.1 Image Stabilization

The images obtained from a lifelog camera occasionally suffer from degraded spatial quality due to the movements of the lifelogger and the lack of focus of the camera, which result in a blurry image. This image degradation could potentially reduce the performance of the spatial feature extractions, which are dependent on the clear boundaries between the objects and the background in the scene. To diminish this problem, we utilize an image stabilization method to reduce the noise and blur and produce a clean image to use at the feature extraction stage.

For this purpose, we use the multi-stage image stabilization method [31], which has state-of-the-art performance in real-world datasets for various restoration tasks such as deblurring and denoising. The method uses multiple stages of encoder-decoders where their outputs are fused together employing a cross-stage supervised attention mechanism. This deblurring process is especially helpful in scenes where the motion blur is the dominant type and the direction of the motion is uniform. We use the official implementation of the method provided by the authors[2] for deblurring the images in the LSC dataset prior to the feature extraction. Example outputs of the method are shown in Figure 3 where the amount of motion blur on the food caused by the movement of the lifelogger is reduced.

Figure 3: The result of image stabilization on a sample from the dataset. The images on the left are the original and on the right are the result of deblurring process.

3.2 Map-based Queries and Result Presentation

Spatial queries in Cineast are executed as nearest-neighbour search using the *haversine distance* on a two-dimensional vector of longitude and latitude, resulting in an approximation of the great-circle distance of any two given points [2]. Given a distance d between two points and a divisor x, we use a hyperbolic correspondence function to calculate the score s of a segment as follows:

$$s = clamp(0, \frac{1}{(1 + \frac{d}{x})}, 1) \tag{1}$$

In the context of same-city queries, as used in [26], we found empirically that a maximal distance of $1000m$ results in reasonable ranks.

As described in Section 4, in the context of LSC, the spatial information is provided as metadata and imported into Cottontail DB. Geo-based queries can be performed in two ways: First, by specifying a circle on a map, allowing the user to express their uncertainty about a location, and second, by entering semantic names provided in the data set (e.g. "home", "work"). The coordinates of these locations are based on the provided metadata. Temporal context can be expressed similarly to all other modalities. Figure 4 shows how the map-based query modality looks like.

For LSC 2021, we plan to update vitrivr-ng to include a map-based result view in addition to the existing views.

[2]https://github.com/swz30/MPRNet

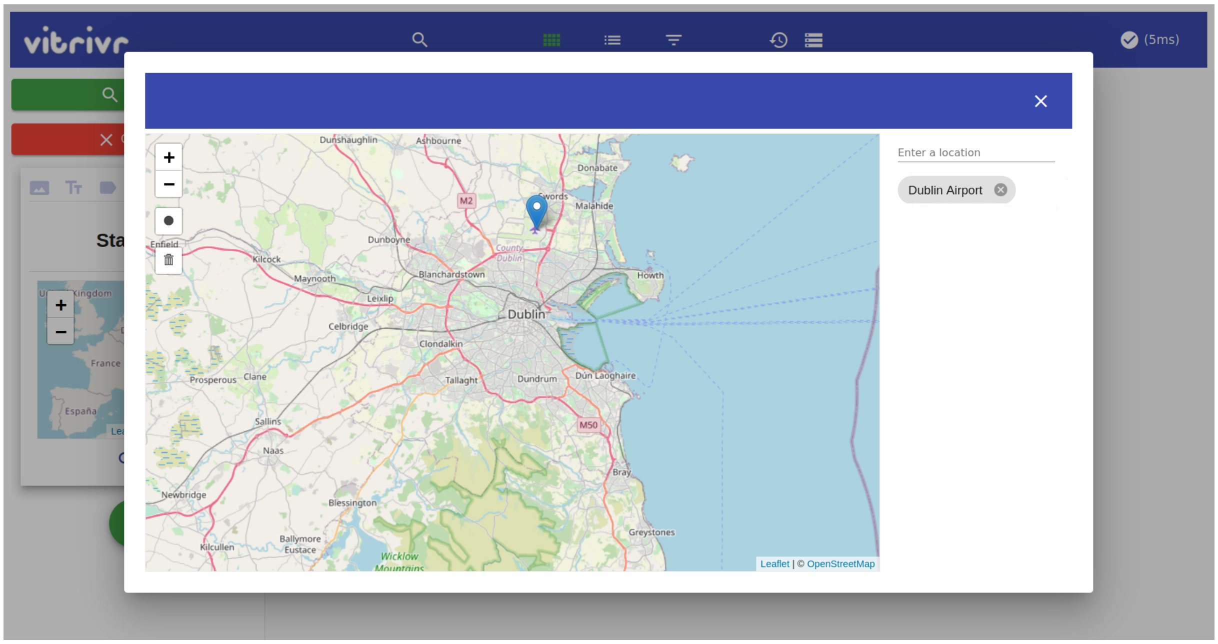

Figure 4: Illustrative screenshot of the map-based query by semantic location.

4 DATASET PREPARATION

In comparison to the 2020 installment of the LSC, the data set changed only slightly and as a consequence, we applied the same data preparation steps as discussed in [9]. An overview of the data set can be found in [8]. In order to support the recent additions to the vitrivr system, as described in Section 3, we enhanced the data slightly: Spatial metadata is imported as a two-dimensional vector consisting of longitude and latitude and transformed from simple, table-like metadata, into feature data usable for similarity queries. Additionally, we are using the Microsoft Cognitive Service Annotations provided by the Myscéal [29] Team. We treated them as an expansion to the provided visual concepts which will be usable for input in Boolean- and *tag*-queries in vitrivr as introduced in [18]. For instance, this enables users to query for "all images with the concept *person*".

5 RELATED WORK

The Lifelog Search Challenge brings diverse multimedia retrieval systems together. Some of last year's LSC participants, such as Exquisitor [28] and Li et al. [15], allow the user to specify whether they find particular results relevant or not. The system then uses this information to improve the result set.

Exquisitor is an interactive learning system, with users deciding which images of a suggested collection match the description on a given LSC task. The user either accepts or declines an image, and the system tries to provide new results based on user feedback. In the system of Li et al., users can exclude similar images from the result set.

VIRET [13], vitrivr [9], and Myscéal [29] offer temporal queries to express an information need that spans multiple, consecutive shots, with vitrivr and Myscéal offering the ability to indicate in which time interval the described scenes happened. Many systems provide the functionality of filtering by day and hour, meanwhile others provide temporal context by showing images and respective shots taken before and after [14, 17, 28, 30].

Most of the systems offer querying by location. For our work the systems Myscéal [29], LifeSeeker [14], and the work of Chu et al. [3] are particularly interesting, since they provide a visualization of spatial context with the help of a map in various different ways:

Myscéal. In Myscéal, the user can query a location by drawing a rectangle on a map. Furthermore, the user does not have to explicitly query a location, since a default map is provided. If the user enters a text query, the map shows the locations of the result images. When the user hovers over the locations on the map, the respective images in the result set are highlighted – or vice versa.

LifeSeeker. LifeSeeker represents spatial context by using a directed graph in which a vertex depicts a location and an edge the chronological order of visited places. There are three granularity levels of the graph: On the first level, the vertices depict, for example, a place in another country, the second level shows more detailed locations (e.g., home, work, airport etc.) and the third level indicates, for example, in which room of the lifelogger's home an image has been taken. The user can then click on a vertex and associated images will be presented. In addition to that, the system recognizes and suggests movement patterns of the lifelogger.

Chu et al. The system of Chu et al. also embeds a map but with the purpose of helping the user to recall all locations that have been visited by the lifelogger, where the visited cities are marked with a pin. Since the focus lies on guiding the user, it is at currently not possible to create a query by indicating the desired location on the map. However, it is possible to query a location by entering a place as keyword.

A completely different approach in user interaction is taken by Voxento [1]. In Voxento, a user can create, delete and execute queries by speaking. The user can then manually edit the text which the system extracted from the user's spoken sentence. The results are updated in real-time while the user is adding more information. The system is able to respond and tell the size of the result set or the metadata of a selected image.

6 CONCLUSION

In this paper, we discussed the 2021 participation of the vitrivr system in the LSC. vitrivr uses a combination of query modalities, such as concept-based and OCR text retrieval, color and semantic sketches, and traditional Boolean queries. For this iteration of the LSC, we have added a new map-based query mode and result presentation. Moreover we have also integrated image stabilization techniques into vitrivr. These techniques can be applied to images prior to feature extraction with the objective to increase the quality of the image features and thus to improve the entire retrieval approach for lifelogging data.

ACKNOWLEDGMENTS

This work was partly supported by the Swiss National Science Foundation (project "Polypheny-DB: Cost- and Workload-aware Adaptive Data Management", contract no. 200021_172763) and the Hasler Foundation (project "City-Stories: Spatio-temporal Search over Crowdsourced Content", contract no. 17055).

REFERENCES

[1] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2020. Voxento: A Prototype Voice-controlled Interactive Search Engine for Lifelogs. In *Proceedings of the Third Annual Workshop on the Lifelog Search Challenge (LSC '20)*. ACM.

[2] Lukas Beck and Heiko Schuldt. 2016. City-Stories: A Spatio-Temporal Mobile Multimedia Search System. In *IEEE International Symposium on Multimedia (ISM)*.

[3] Tai-Te Chu, Chia-Chun Chang, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Multimodal Retrieval through Relations between Subjects and Objects in Lifelog Images. In *Proceedings of the Third Annual Workshop on the Lifelog Search Challenge (LSC '20)*. ACM.

[4] Ralph Gasser, Luca Rossetto, Silvan Heller, and Heiko Schuldt. 2020. Cottontail DB: an open source database system for multimedia retrieval and analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*.

[5] Ralph Gasser, Luca Rossetto, Silvan Heller, and Heiko Schuldt. 2021. Multimedia Retrieval and Analysis with Cottontail DB. *ACM SIGMM Records, Open Source Column* (2021).

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

[7] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Dang Nguyen, Duc Tien, Michael Riegler, Luca Piras, Minh-Triet Tran, Jakub Lokoč, and Wolfgang Hürst. 2019. Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC2018). *ITE Transactions on Media Technology and Applications* 7, 2 (2019).

[8] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2021. Introduction to the Fourth Annual Lifelog Search Challenge, LSC'21. In *Proc. International Conference on Multimedia Retrieval (ICMR'21)*. ACM, Taipei, Taiwan.

[9] Silvan Heller, Mahnaz Amiri Parian, Ralph Gasser, Loris Sauter, and Heiko Schuldt. 2020. Interactive lifelog retrieval with vitrivr. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*.

[10] Silvan Heller, Ralph Gasser, Cristina Illi, Maurizio Pasquinelli, Loris Sauter, Florian Spiess, and Heiko Schuldt. 2021. Towards explainable interactive multi-modal video retrieval with vitrivr. In *International Conference on Multimedia Modeling*.

[11] Silvan Heller, Mahnaz Parian, Maurizio Pasquinelli, and Heiko Schuldt. 2020. Vitrivr-Explore: Guided Multimedia Collection Exploration for Ad-hoc Video Search. In *International Conference on Similarity Search and Applications*.

[12] Silvan Heller, Loris Sauter, Heiko Schuldt, and Luca Rossetto. 2020. Multi-Stage Queries and Temporal Scoring in Vitrivr. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*.

[13] Gregor Kovalčík, Vít Škrhák, Tomáš Souček, and Jakub Lokoč. 2020. VIRET Tool with Advanced Visual Browsing and Feedback. In *Proceedings of the Third Annual Workshop on the Lifelog Search Challenge (LSC '20)*. ACM.

[14] Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Thanh-An Nguyen, Hai-Dang Nguyen, Liting Zhou, Graham Healy, and Cathal Gurrin. 2020. LifeSeeker 2.0 : Interactive Lifelog Search Engine at LSC 2020. In *Proceedings of the Third Annual Workshop on the Lifelog Search Challenge (LSC '20)*. ACM.

[15] Jiayu Li, Min Zhang, Weizhi Ma, Yiqun Liu, and Shaoping Ma. 2020. A Multi-level Interactive Lifelog Search Engine with User Feedback. In *Proceedings of the Third Annual Workshop on the Lifelog Search Challenge (LSC '20)*. ACM.

[16] Jakub Lokoč, Gregor Kovalčík, Bernd Münzer, Klaus Schöffmann, Werner Bailer, Ralph Gasser, Stefanos Vrochidis, Phuong Anh Nguyen, Sitapa Rujikietgumjorn, and Kai Uwe Barthel. 2019. Interactive search or sequential browsing? A detailed analysis of the video browser showdown 2018. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1 (2019).

[17] František Mejzlík, Patrik Veselý, Miroslav Kratochvíl, Tomáš Souček, and Jakub Lokoč. 2020. SOMHunter for Lifelog Search. In *Proceedings of the Third Annual Workshop on the Lifelog Search Challenge (LSC '20)*. ACM.

[18] Luca Rossetto, Ralph Gasser, Silvan Heller, Mahnaz Amiri Parian, and Heiko Schuldt. 2019. Retrieval of structured and unstructured data with vitrivr. In *Proceedings of the Second Annual Workshop on Lifelog Search Challenge*.

[19] Luca Rossetto, Ralph Gasser, Silvan Heller, Mahnaz Parian-Scherb, Loris Sauter, Florian Spiess, Heiko Schuldt, Ladislav Peska, Tomas Soucek, Miroslav Kratochvil, et al. 2021. On the User-centric Comparative Remote Evaluation of Interactive Video Search Systems. *IEEE MultiMedia* (2021).

[20] Luca Rossetto, Ralph Gasser, Jakub Lokoc, Werner Bailer, Klaus Schoeffmann, Bernd Muenzer, Tomas Soucek, Phuong Anh Nguyen, Paolo Bolettieri, Andreas Leibetseder, et al. 2020. Interactive video retrieval in the age of deep learning-detailed evaluation of vbs 2019. *IEEE Transactions on Multimedia* (2020).

[21] Luca Rossetto, Ralph Gasser, Loris Sauter, Abraham Bernstein, and Heiko Schuldt. 2021. A System for Interactive Multimedia Retrieval Evaluations. In *International Conference on Multimedia Modeling*.

[22] Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. 2014. Cineast: a multi-feature sketch-based video retrieval engine. In *IEEE International Symposium on Multimedia, ISM*. IEEE Computer Society.

[23] Luca Rossetto, Ivan Giangreco, Claudiu Tanase, and Heiko Schuldt. 2016. vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections. In *Proceedings of the 24th ACM international conference on Multimedia*.

[24] Luca Rossetto, Mahnaz Amiri Parian, Ralph Gasser, Ivan Giangreco, Silvan Heller, and Heiko Schuldt. 2019. Deep learning-based concept detection in vitrivr. In *International Conference on Multimedia Modeling*.

[25] Loris Sauter, Mahnaz Amiri Parian, Ralph Gasser, Silvan Heller, Luca Rossetto, and Heiko Schuldt. 2020. Combining boolean and multimedia retrieval in vitrivr for large-scale video search. In *International Conference on Multimedia Modeling*.

[26] Loris Sauter, Luca Rossetto, and Heiko Schuldt. 2018. Exploring Cultural Heritage in Augmented Reality with GoFind!. In *IEEE International Conference on Artificial Intelligence and Virtual Reality, AIVR*.

[27] Klaus Schoeffmann. 2019. Video Browser Showdown 2012-2019: A Review. In *International Conference on Content-Based Multimedia Indexing, CBMI*.

[28] Omar Shahbaz Khan, Mathias Dybkjær Larsen, Liam Alex Sonto Poulsen, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, Dennis Koelma, and Marcel Worring. 2020. Exquisitor at the Lifelog Search Challenge 2020. In *Proceedings of the Third Annual Workshop on the Lifelog Search Challenge (LSC '20)*. ACM.

[29] Ly-Duyen Tran, Manh-Duy Nguyen, Nguyen Thanh Binh, Hyowon Lee, and Cathal Gurrin. 2020. Myscéal: An Experimental Interactive Lifelog Retrieval System for LSC'20. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*.

[30] Minh-Triet Tran, Thanh-An Nguyen, Quoc-Cuong Tran, Mai-Khiem Tran, Khanh Nguyen, Van-Tu Ninh, Tu-Khiem Le, Hoang-Phuc Trang-Trung, Hoang-Anh Le, Hai-Dang Nguyen, Trong-Le Do, Viet-Khoa Vo-Ho, and Cathal Gurrin. 2020. FIRST - Flexible Interactive Retrieval SysTem for Visual Lifelog Exploration at LSC 2020. In *Proceedings of the Third Annual Workshop on the Lifelog Search Challenge (LSC '20)*. ACM.

[31] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. 2021. Multi-Stage Progressive Image Restoration. *CoRR* abs/2102.02808 (2021).