Exploring Intuitive Lifelog Retrieval and Interaction Modes in Virtual Reality with vitrivr-VR

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

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

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

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

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

Milan van Zanten  
University of Basel  
Basel, Switzerland  
milan.vanzanten@unibas.ch

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

ABSTRACT
The multimodal nature of lifelog data collections poses unique challenges for multimedia management and retrieval systems. The Lifelog Search Challenge (LSC) offers an annual evaluation platform for such interactive retrieval systems. They compete against one another in finding items of interest within a set time frame.

In this paper, we present the multimedia retrieval system vitrivr-VR, the latest addition to the vitrivr stack, which participated in the LSC in recent years. vitrivr-VR leverages the 3D space in virtual reality (VR) to offer novel retrieval and user interaction models, which we describe with a special focus on design decisions taken for the participation in the LSC.

CCS CONCEPTS
- Information systems → Image search; Search interfaces; Query representation. → Human-centered computing → Virtual reality.

KEYWORDS
Lifelog Search Challenge, Virtual Reality, Interactive Lifelog Retrieval

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

1 INTRODUCTION
A lifelog is a record of different aspects of one’s personal activities throughout the day at varying levels of detail. With smartphones and wearable devices becoming more prolific, the act of lifelogging has become more accessible to a growing audience and the information that can be gathered is expanding both in terms of volume and variety. The data typically found in a lifelog can be anything from highly structured information, such as metrics about one’s physical condition (e.g., blood pressure, pulse) or metadata about ones location (e.g., coordinates, speed), to highly unstructured information, such as images or video sequences of random encounters during the day. Some lifeloggers even use wearable cameras that take pictures at regular intervals to create an extensive log of their daily routine.

There are many different motivations to keep a lifelog, ranging from a hobby of technology enthusiasts to concrete applications in personal healthcare and medicine to keeping a record of certain activities in professional life. Irrespective of the motivation, lifelogging poses very interesting research questions in regards to the storage, management, and search in such multi-modal data collections, especially given the huge amounts of information one can accumulate over a lifetime. It is therefore not surprising that a dedicated field of research has emerged around the subject of lifelogging [10] and retrieval in lifelog data collections [9].

The Lifelog Search Challenge (LSC) [11] is an annual event co-located with the International Conference on Multimedia Retrieval (ICMR). The LSC offers a platform for testing and comparing information retrieval systems on lifelog data. The competitive setting involves solving a selection of retrieval tasks within a set time limit. The 2021 installment of the LSC [11] operates on the same dataset used during the LSC 2020 [9]. It covers 115 individual days between February 23, 2015 and May 05, 2018. The data contains a total of 193’911 images from two different perspectives in a 1024px × 768px resolution as well as the associated metadata such as local time, geo references, sensor readings, and other information.
In this paper we present vitrivr-VR, a first-time participant in the LSC. This newest addition to the vitrivr stack [26], which has participated in the LSC several times already, offers a novel user interaction and retrieval model that leverages the 3D space offered by virtual reality (VR). This paper provides an overview of the vitrivr-VR system participating in the LSC 2021, with a focus on the aspects and design decisions most relevant for the LSC.

The remainder of this paper is structured as follows: Section 2 briefly surveys related work. Section 3 gives an overview of vitrivr-VR’s system architecture and describes the retrieval modalities and the novel user interaction model employed, and Section 4 concludes and offers an outlook.

2 RELATED WORK

Lifelogging as a dedicated area of research has already been proposed and described in 2014 [10] with a focus on its information retrieval aspects. The Lifelog Search Challenge (LSC) 2021 [11] is the fourth installment of the LSC and offers an annual platform for interactive lifelog retrieval systems to compare and compete. vitrivr is taking part for the third time [14], with previous participation in 2019 and 2020. In both previous cases, a conventional, web-based user interface [8] has been used. The main contribution in 2019 was extending vitrivr to support pure Boolean retrieval queries, which are a requirement for the LSC [20], whereas in 2020, we added support for temporal scoring, staged querying and deep learning based concept detection [12]. One of the main takeaways of these two iterations is that an effective user interface for result exploration is at least as important as an expressive retrieval model. This is also a lesson learned from analyses of the Video Browser Showdown (VBS) [18, 22], where vitrivr is an active participant [13, 27, 28], vitrivr has experimented with exploratory user interfaces in the past [15].

The idea of using VR as a means of user interaction in lifelog retrieval is not new. In fact, the very first winner of the LSC [4] – the Virtual Reality Lifelog Explorer (VRLE) – featured a VR UI and has been a participant again in 2020 [5]. While one of the first and most effective VR lifelog exploration systems, VRLE is a rather conservative VR adaptation of a conventional 2D user interface and even though different modes of user interaction were explored in [2], the query menu and horizontally scrolling result display are very similar to conventional UIs and do not take full advantage of the virtual space.

In strong contrast to VRLE, the lifelog exploration system described in [17] forges most conventional query modes and user interface elements, providing only a VR map to explore the lifelog data through their associated geospatial information. Although this results in an intuitive way to explore the spatial aspects of the lifelog data in a way not possible using conventional user interfaces, the restriction to one aspect of the data severely limits the system’s ability to take advantage of much of the information provided in lifelog search tasks, such as temporal and visual semantic information.

While the current research corpus on VR user interfaces for lifelog retrieval and multimedia retrieval in general is not conclusive, a systematic comparison of VRLE to a classical 2D system [3] indicates that there is no clear advantage of using a 2D system, despite the mode of interaction being far more common. This shows that VR-based multimedia retrieval is a largely unexplored avenue for research. Therefore, in our work, we aim to combine the most effective, known methods from conventional retrieval interfaces with novel, purpose-built approaches that take advantage of virtual reality environments. First steps in that direction have already led to an experimental Query-by-Sculpting interface for retrieval in 3D model collections [1] and a VR based participation in the VBS [30].

3 SYSTEM OVERVIEW

vitrivr is a general purpose, open-source multimedia retrieval stack with support for images, audio, video, and 3D model collections [7], and both content-based similarity search as well as Boolean retrieval in metadata. The vitrivr-VR user interface, described herein, can be seen as the latest addition to the stack and uses all the default programming interfaces offered by vitrivr. Figure 1 illustrates the full stack of vitrivr and vitrivr-VR, with its three major components:

(i) Cottontail DB [6] is the database management system that stores all relevant information such as feature representations extracted from the multimedia data as well as any metadata. Furthermore, Cottontail DB facilitates execution of Boolean and nearest neighbor search (NNS) queries used in multimedia retrieval, (ii) Cineast [24, 25] is the retrieval engine at the core of the stack and handles query processing and feature transformation both during offline feature extraction from media collections as well as during online retrieval. Additionally, it takes care of query orchestration, score fusion and result set re-construction. An overview of its retrieval model can be found in [16], and (iii) vitrivr-VR is the experimental, VR-based user interface presented in this paper, which offers all necessary functionality for query formulation and result exploration.

For the sake of completeness, the classical, web-based user interface, vitrivr-ng, is also depicted in Figure 1. vitrivr-VR sees itself as a prototypical extension to the stack in an attempt to investigate multimedia retrieval in virtual reality. As such, it complements vitrivr-ng and does not aim to replace it. vitrivr-VR is developed in Unity3 with C# for the HTC Vive Pro and Valve Index, and communicates with Cineast using the RESTful API provided through its OpenAPI specifications.

vitrivr-VR interfaces with VR hardware through the Unity OpenXR plugin3. We have chosen to target OpenXR as VR integration offers more platform-specific options, such as the OpenVR plugin, even though this limits hardware functionality to the subset exposed through the OpenXR interface, for two reasons. Firstly, targeting OpenXR allows vitrivr-VR to be developed mostly XR-backend agnostic, allowing it to be easily and quickly deployed to hardware not available during development or not supported by SteamVR. Secondly, the Unity OpenXR bindings are completely interaction system independent, which makes it easier to swap interaction systems or experiment with custom interactions. As interaction system for vitrivr-VR, we use a mixture of the still in early development Unity XR Interaction Toolkit4 and our own custom interaction

---

3 https://docs.unity3d.com/Packages/com.unity.xr.interaction.toolkit@latest
4 https://docs.unity3d.com/Packages/com.unity.xr.interaction.toolkit@latest
3.1 Retrieval Modalities

vitrivr-VR makes use of three main retrieval modalities: concept-based, text-based, and spatio-temporal retrieval. These modalities are a subset of the ones also available to the web-based interface vitrivr-ng and were selected for their effectiveness in lifelog retrieval tasks based on our experiences from previous instances of the LSC.

Concept-based retrieval is enabled through Boolean search in Cottontail DB and performed on the concepts bundled with the LSC dataset, which are complemented by automatically extracted labels. Concepts may include prominent objects or the lifelogger’s current activity and are often subject of lifelog queries and therefore play an important role in lifelog retrieval.

Spatial (e.g., “I was in Oslo.”) and temporal information (e.g., “It was morning.”) also played a very important role in a lot of LSC tasks in recent years, which is why we included these modalities as well. Temporal retrieval is again powered by classical Boolean search, whereas spatial queries are realized through nearest neighbor search using spherical coordinates and the Haversine distance.

Full-text search is used mainly for searching the automatically generated image captions, as was used by vitrivr in previous iterations of the LSC. In addition, vitrivr-VR also leverages a text-based retrieval feature based on a visual-text co-embedding. Capable of capturing a more detailed semantic representation than concept-tags, while being more flexible to different formulations than fixed scene captions, the effectiveness of visual-text co-embeddings was demonstrated in [19] and in an in-depth comparison of vitrivr and another system using similar embeddings [21].

Similarity search features, e.g., based on example images or color sketches, were not included in this iteration, since their use for the LSC is expected to be limited as a result of the textual nature of the task descriptions.

3.2 User Interaction in Virtual Reality

The user interaction in vitrivr-VR has been designed with interactivity and immersion at its core. Great care has been taken to tailor both the interactive retrieval process, which encompasses all interaction in vitrivr-VR from the initial query to query refinement and result exploration, as well as the interactions used to specify each aspect of a query for VR. In both of these areas, our aim was to adapt the most effective methods from systems designed for conventional interfaces to VR, and to combine them with new methods developed entirely with VR in mind.

3.2.1 Interactive Retrieval Process.

The interactive retrieval process in vitrivr-VR is designed to allow iterative query refinement and query management. The most important components of this interactive retrieval process are the individual queries, the user-placeable result detail views and the query management.

The retrieval process starts with an initial query consisting of text-, concept- and spatio-temporal query terms. As the user browses the results and new information may have been added to the retrieval prompt, the user can alter the query and retrieve a new set of results. Not only does vitrivr-VR store the results of user issued queries, it also retains the user’s browsing position when a new query is issued to allow users to quickly return to exploring
previous result sets in case the refined query does not yield more promising results.

To make use of the virtual space during the interactive retrieval process, result detail views may be pulled out of a result set and placed anywhere in space. These pinned results also remain when switching between query results and thereby allow results to be compared and considered across queries before submission. The result detail views also expose further information about results of interest, such as associated concepts and metadata, which may be used to further refine the query.

3.2.2 Text-based Query Formulation. Text-based queries and also concept-based queries have proven to be very important in a lifelog search setting and both rely on textual input by the user. As primary text input method, vitrivr-VR uses speech-to-text, which allows fast, hands-free input of text in VR. Speech-to-text makes it easy to specify long text-based queries that are difficult and tedious to type without a physical keyboard. This makes speech-to-text especially well suited for full-text search on the automatically generated image captions and as input to the visual-text co-embedding. While individual keywords work well for concept-based search, the visual-text co-embedding is able to benefit from the additional semantic information contained in full sentences captured through speech-to-text.

In addition to speech-to-text, an optional contact-based virtual keyboard can be used to make small corrections and input hard-to-pronounce or non-English words. This contact-based keyboard is most useful for search of visible text extracted through OCR, as these often contain building and place names that may contain very unique spelling difficult to specify through speech-to-text.

3.2.3 Spatial Query Formulation. To make use of the spatial information contained in the lifelog data, vitrivr-VR allows the formulation of spatial queries through a map-based interface. The map-based query formulation interface consists of two parts as seen in Figure 2: a flat, zoom- and scrollable detail-map and a globe mini-map. While the globe mini-map can be used to roughly pinpoint a location, the detail-map can be used to specify query locations with high precision by adjusting the map’s level of zoom to the required level of detail. The two maps are connected such that approximate locations of interest can be selected in the globe mini-map to be displayed and viewed on the detail-map, from which query locations can then be selected with high precision.

The interaction modes of the detail-map are inspired by existing touchscreen-based map interactions and those used with physical maps. Scroll interactions are performed by grabbing and dragging the map with either hand, while the level of zoom is adjusted by grabbing the map with both hands and moving them closer together to zoom out and farther away to zoom in, similar to the pinch gesture used in touch interfaces. Specifying query locations is done by placing virtual pins, such as the one seen in the top right of Figure 2, at the desired locations. To avoid the need to repeatedly change the map’s zoom level when scrolling large distances, the globe mini-map allows users to move the center of the detail-map quickly to a rough location anywhere on the globe. Interacting with the globe mini-map is analogous to interacting with a physical globe, allowing it to be grabbed, moved and rotated to be easily viewed from any angle.

The map data used by the spatial query formulation interface is loaded through a modified version of the Mapbox Unity Package and includes OpenStreetMap data.

3.2.4 Result Exploration. Query results are arranged cylindrically around the user as postcard-sized images, as seen in Figure 3. The view retains the grid structure intuitive to users of conventional interfaces, while making use of the additional freedoms of VR by surrounding the user. Results can be scrolled horizontally, revealing further results replacing those already viewed. While this method

---

6https://www.mapbox.com/unity/
7https://www.openstreetmap.org
of result display is reminiscent of result displays in conventional interfaces [29] and does not make full use of virtual space by itself, this single-layered approach avoids result images occluding each other while allowing a large number of results to be browsed simply by looking around the virtual space.

Individual result items can be ‘popped’ out of the display and shown in a detailed view that also gives access to metadata and detected concepts. These result item detail views can be moved and positioned around the virtual space as reference or for later consideration. This allows individual results to be compared across queries, neighboring segments to be viewed and even the segment to be submitted to the competition system [23] for evaluation at any time in a competition setting.

In the lifelog setting, a piece of data is rarely independent and usually only appears in the context of a series of data. To contextualize query results within the data-series from which they originate, vitrivr-VR offers a sequence view for each result item, accessible from their detail view. An example of a sequence view is shown in Figure 4.

This sequence view shows the selected result item and a number of its sequence neighbors as slices inside a box in virtual space. Individual slices from this sequence can be inspected by hovering over them with a hand. By pulling the trigger on the selecting hand, the detailed view of the selected item can be accessed. The handle at the front of the view can be pulled toward the user like a drawer handle to elongate the sequence view and increase the spacing between the sequence items, making browsing easier. This sequence view is very space-efficient and allows a sequence of lifelog images to be browsed more quickly and intuitively than through conventional flat interfaces.

4 CONCLUSION

In this paper, we presented vitrivr-VR, a virtual reality multimedia retrieval system. vitrivr-VR features a VR interface designed to combine the most effective methods from conventional user interfaces translated into the virtual space with interaction methods only possible in VR. We expect vitrivr-VR to be competitive in a lifelog search competition setting and to gain valuable insights into both the application of methods from conventional interfaces in VR as well as VR specific interaction methods.

While we take a big step in the direction of interactive multimedia retrieval designed for VR with vitrivr-VR, there is still a lot of untrod ground to investigate. In terms of both result display and exploration, there are still many aspects to be explored, such as display methods that make efficient use of virtual space and the balance between virtual space usage and visual overload. Furthermore, the influence of the virtual environment during tasks such as multimedia search and exploration has yet to be investigated.

ACKNOWLEDGMENTS

This work was partly supported by the Swiss National Science Foundation (project “Participatory Knowledge Practices in Analogue and Digital Image Archives”, contract no. CRSII5_193788) and the Hasler Foundation (project “City-Stories: Spatio-temporal Search over Crowdsourced Content”, contract no. 17055).

REFERENCES


