Competitive Interactive Video Retrieval in Virtual Reality with vitrivr-VR

Florian Spiess[1][0000−0002−3396−1516], Ralph Gasser[1][0000−0002−3016−1396], Silvan Heller[1][0000−0001−5386−330X], Luca Rossetto[2][0000−0002−5389−9465], Loris Sauter[1][0000−0001−8046−0362], and Heiko Schuldt[1][0000−0001−9865−6371]

[1] Department of Mathematics and Computer Science
University of Basel, Basel, Switzerland
{firstname.lastname}@unibas.ch
[2] Department of Informatics, University of Zurich, Zurich, Switzerland
rossetto@ifi.uzh.ch

Abstract. Virtual Reality (VR) has emerged and developed as a new modality to interact with multimedia data. In this paper, we present vitrivr-VR, a prototype of an interactive multimedia retrieval system in VR based on the open source full-stack multimedia retrieval system vitrivr. We have implemented query formulation tailored to VR: Users can use speech-to-text to search collections via text for concepts, OCR and ASR data as well as entire scene descriptions through a video-text co-embedding feature that embeds sentences and video sequences into the same feature space. Result presentation and relevance feedback in vitrivr-VR leverages the capabilities of virtual spaces.

Keywords: Video Browser Showdown · Virtual Reality · Interactive Video Retrieval

1 Introduction

The Video Browser Showdown (VBS) [14] is an annual interactive video search competition where video retrieval systems are evaluated competitively through a variety of search tasks on large video datasets.

Virtual reality (VR) technology has developed rapidly in recent years, and due to technological advances that have continued to make the hardware required more accessible, affordable, and comfortable, VR capable devices have become much more commonplace. VR has great potential when interacting with multimedia, especially in the context of retrieval, due to the ability of users to be immersed in a virtual space, which provides a near limitless presentation area. Additionally, hand and head tracking directly translates real world movement into movement within the virtual space, making spatial orientation intuitive.

We leverage components of the existing vitrivr system [16], which has participated at VBS for a number of years [4,18], and have built a new system, vitrivr-VR, with a virtual reality based interface to tackle the problem of efficient and intuitive interactive video retrieval in large video collections by exploiting

the potential of multimedia retrieval in VR. vitrivr-VR has voice input for initial queries and uses a number of query refinement and relevance feedback tools engineered to enable intuitive and efficient video search in VR.

In the remainder of this paper we contextualize our work with respect to existing research in Section 2, present an overview of the vitrivr-VR architecture in Section 3, describe the query mechanisms used in Section 4, outline the interactive retrieval process in Section 5, and conclude in Section 6.

2 VR Multimedia Retrieval Interfaces

A few approaches to perform multimedia retrieval and exploration in VR have already been proposed and have shown the potential of this new modality [9], even in competitive settings such as the Lifelog Search Challenge (LSC) [1], where vitrivr has also been a participant in recent years [5]. While these existing systems have shown that multimedia retrieval and exploration is possible in VR, they primarily investigate how traditional 2D interfaces can be transferred into the VR space without significant changes to the mode of interaction and media presentation. In [1], for example, different methods of interacting with traditional 2D user interfaces in VR are explored, however, the user interfaces themselves are not adapted to the virtual space beyond their placement in the 3D space. Although these interfaces are immediately familiar from traditional 2D interfaces, they lose much of their efficiency in VR, especially for complex interactions, such as input of specific text.

Other approaches propose methods to make use of the space and freedom of movement available in VR. [3] introduces a system that enables 3D sketch-based model retrieval in VR through the fine-granular input possible with 3D tracked controllers. The retrieval results are displayed in a traditional 2D scrolling list. A system that explores result presentation in VR is described in [13], which proposes a simple approach to result presentation in virtual spaces by directly mapping feature similarity scores to the three spatial dimensions.

With vitrivr-VR we deliberately attempt to avoid traditional 2D user interfaces and presentation methods designed for 2D displays in favor of investigating the effectiveness of user interfaces designed with virtual spaces in mind.

3 System Overview

As the name suggests, vitrivr-VR shares some components with vitrivr, an open source multimedia retrieval stack.[3] It also uses Cottontail DB [2] as its storage layer and Cineast [15] as its retrieval engine. The presentation layer is a VR environment as opposed to the traditional web interface. Cottontail DB is a column store that allows combining Boolean retrieval and nearest neighbor search. Cineast is responsible for both retrieval and feature extraction, supporting a multitude of features and media types.

[3] https://vitrivr.org

To communicate with Cineast, vitrivr-VR uses the RESTful API provided through the OpenAPI specifications. vitrivr-VR is implemented in Unity[4] using the Unity XR Interaction Toolkit[5] to interface with the HTC Vive Pro headset through OpenVR.

The existing components from the vitrivr stack provide the core multimedia retrieval capabilities [7], while vitrivr-VR enables query formulation, query refinement, relevance feedback and browsing through its VR user interface. vitrivr-VR provides speech-to-text based text input for textual queries, query-by-example (QbE) through previously retrieved results and a number of query refinement and relevance feedback options.

4 Querying Mechanisms

vitrivr offers a multitude of feature descriptors, many of which focusing on various visual aspects of a video. Since the primary query formulation mechanism for vitrivr-VR is speech input which is automatically transcribed into text, we limit ourselves to the features which operate on textual input. These include previously introduced capabilities [17], such as full-text search in automatically transcribed dialog, visible text or automatically generated scene captions as well as tag-based search using a large number of individually detected instances of objects or semantic concepts visible in a scene. In addition, we use a video-text co-embedding mechanism inspired by [11], which is capable of capturing a richer semantic representation when compared to the individual tags while understanding a larger vocabulary than the already used captioning approach. The effectiveness of such methods has recently been demonstrated in [12].

We also make use of relevance feedback for non text-based querying, which is used by many systems in interactive video retrieval [6,8,10] and is also well suited for the result presentation and interaction mechanisms of vitrivr-VR.

5 Interactive Retrieval Process in VR

In vitrivr-VR, using the system to find the figurative needle in the haystack is an interactive process which can be categorised into three phases; *initial query*, *result organisation* and *refinement queries* A simplified interaction flow diagram of the process is shown in Figure 1.

5.1 Initial Query

Recent VBS installments have shown that text-based retrieval, be it for a limited, known set of concepts or automatically generated scene captions, is a successful strategy for an initial query. However, typing words for text-based search is not very efficient in VR since the state of the art in finger-tracking for text input is not yet fast and stable. Hence, vitrivr-VR relies on speech-to-text[6] to formulate

[4] https://unity.com
[5] https://docs.unity3d.com/Packages/com.unity.xr.interaction.toolkit@0.9
[6] https://docs.microsoft.com/en-us/windows/mixed-reality/voice-input-in-unity

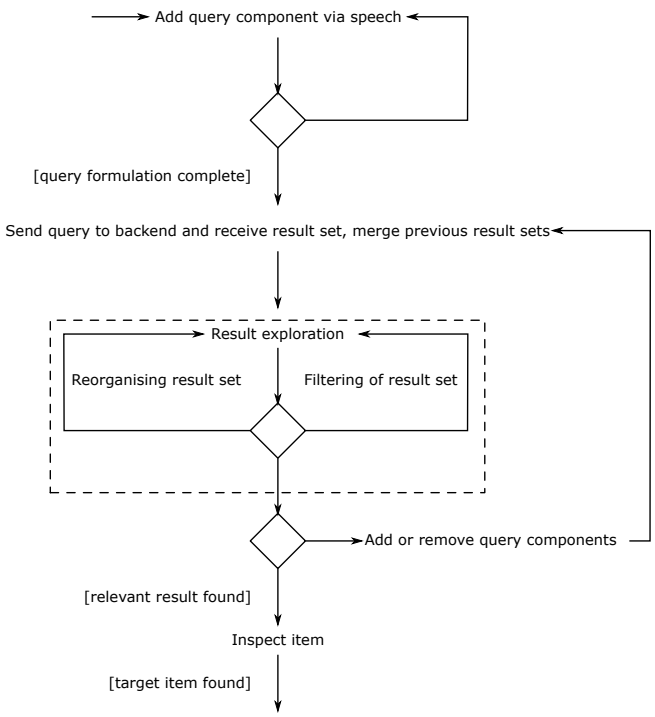

Fig. 1. Interaction flow diagram of vitrivr-VR

the initial query. In particular, vitrivr-VR supports traditional full-text search in OCR, ASR and scene caption data, dedicated concept search, as well as video-text co-embedding search as described in Section 4. The OCR and ASR data is the same as used in [17]. At the end of the initial query, a text-based similarity query is issued.

5.2 Result Organisation

In the second phase, during *result organisation*, users are able to re-arrange the results in order to effectively browse the result collection. One advantage of VR is the virtually unlimited space for result presentation, which we exploit in this phase by initially displaying the result set spherically around the user, where each item's spatial context is indicative of its score. Media items can be pulled from the results display and manually positioned in space. Items 'pinned' in this way will return to their position in the results display once unpinned. A list of concepts sorted by their frequency within the result set is presented in close proximity to the user. Subsequently, users are enabled to spatially cluster the results along certain attributes such as concepts by dragging them from a list and positioning them in space, or remove items from the result set matching a certain filter criterion. The goal of this result organisation is to determine whether the

searched item is already in the result set. Each item can be inspected to view the associated video segment, neighboring segments as well as additional information such as a list of concepts detected within the segment. From this view, items can also be submitted to the competition system. In case the target item is not yet in the result set, the third phase is launched.

5.3 Refinement Queries

From the previous phases, a set of query components produced a result set which was organised and filtered, and is spatially explorable within vitrivr-VR. In case a user did not yet succeed in finding the target item, our system provides multiple means to facilitate query refinement and additional queries: (i) query-by-example for result set expansion or reduction, (ii) more-like-this queries, both visually and semantically, and (iii) relevance feedback. Result sets of these new queries might be seamlessly merged into the pre-existing result set without user interaction. Alternatively, taking advantage of the available space in VR, results of additional queries can be presented spatially separated from the original query in order to give users more control over the merging of the two result sets. Additionally, a query and its result set can be stashed away temporarily or discarded permanently, so the user can focus on other queries.

6 Conclusion

In this paper, we introduced vitrivr-VR, a virtual reality multimedia retrieval system integrated into the open source vitrivr stack. vitrivr-VR provides a VR interface for intuitive query formulation and results presentation in a virtual space. We expect video retrieval in VR to be a competitive, intuitive and, user-friendly alternative to traditional 2D interfaces and for our VBS participation to provide us with valuable insights into the effectiveness of VR based systems in competitive interactive video retrieval.

Acknowledgements

This work was partly supported by the Hasler Foundation in the context of the project City-Stories (contract no. 17055).

References

1. Duane, A., Þór Jónsson, B., Gurrin, C.: Vrle: Lifelog interaction prototype in virtual reality: Lifelog search challenge at acm icmr 2020. In: Proceedings of the Third Annual Workshop on Lifelog Search Challenge (2020)
2. Gasser, R., Rossetto, L., Heller, S., Schuldt, H.: Cottontail db: An open source database system for multimedia retrieval and analysis. In: Proceedings of the 28th ACM International Conference on Multimedia (2020)

3. Giunchi, D., James, S., Steed, A.: 3d sketching for interactive model retrieval in virtual reality. In: Proceedings of the Joint Symposium on Computational Aesthetics and Sketch-Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering (2018)
4. Heller, S., Gasser, R., Illi, C., Pasquinelli, M., Sauter, L., Spiess, F., Schuldt, H.: Towards explainable interactive multi-modal video retrieval with vitrivr. In: International Conference on Multimedia Modeling MMM (2021)
5. Heller, S., Parian, M.A., Gasser, R., Sauter, L., Schuldt, H.: Interactive lifelog retrieval with vitrivr. In: Proceedings of the Third ACM Workshop on Lifelog Search Challenge (2020)
6. Heller, S., Parian-Scherb, M., Pasquinelli, M., Schuldt, H.: vitrivr-explore: Guided multimedia collection exploration for ad-hoc video search. In: International Conference on Similarity Search and Applications (2020)
7. Heller, S., Sauter, L., Schuldt, H., Rossetto, L.: Multi-stage queries and temporal scoring in vitrivr. In: IEEE International Conference on Multimedia & Expo Workshops (2020)
8. Jónsson, B.Þ., Khan, O.S., Koelma, D.C., Rudinac, S., Worring, M., Zahálka, J.: Exquisitor at the video browser showdown 2020. In: International Conference on Multimedia Modeling (2020)
9. Khanwalkar, S., Balakrishna, S., Jain, R.: Exploration of large image corpuses in virtual reality. In: Proceedings of the 24th ACM international conference on Multimedia (2016)
10. Kratochvíl, M., Veselỳ, P., Mejzlík, F., Lokoč, J.: Som-hunter: Video browsing with relevance-to-som feedback loop. In: International Conference on Multimedia Modeling (2020)
11. Li, X., Xu, C., Yang, G., Chen, Z., Dong, J.: W2vv++ fully deep learning for ad-hoc video search. In: Proceedings of the 27th ACM International Conference on Multimedia (2019)
12. Lokoč, J., Souček, T., Veselý, P., Mejzlík, F., Ji, J., Xu, C., Li, X.: A w2vv++ case study with automated and interactive text-to-video retrieval. In: ACM Multimedia (2020)
13. Nakazato, M., Huang, T.S.: 3d mars: Immersive virtual reality for content-based image retrieval. In: IEEE International Conference on Multimedia and Expo (2001)
14. Rossetto, L., Gasser, R., Lokoc, J., Bailer, W., Schoeffmann, K., Muenzer, B., Soucek, T., Nguyen, P.A., Bolettieri, P., Leibetseder, A., et al.: Interactive video retrieval in the age of deep learning-detailed evaluation of vbs 2019. IEEE Transactions on Multimedia (2020)
15. Rossetto, L., Giangreco, I., Heller, S., Tănase, C., Schuldt, H.: Searching in video collections using sketches and sample images–the cineast system. In: International Conference on Multimedia Modeling (2016)
16. Rossetto, L., Giangreco, I., Tanase, C., Schuldt, H.: Vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections. In: Proceedings of the 24th ACM international conference on Multimedia (2016)
17. Rossetto, L., Parian, M.A., Gasser, R., Giangreco, I., Heller, S., Schuldt, H.: Deep learning-based concept detection in vitrivr. In: International Conference on Multimedia Modeling (2019)
18. Sauter, L., Parian, M.A., Gasser, R., Heller, S., Rossetto, L., Schuldt, H.: Combining boolean and multimedia retrieval in vitrivr for large-scale video search. In: International Conference on Multimedia Modeling (2020)