Simplified Video Retrieval in Virtual Reality with vitrivr-VR

Florian Spiess[1][0000−0002−3396−1516], Luca Rossetto[2][0000−0002−5389−9465], 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] Dublin City University, Dublin, Ireland
luca.rossetto@dcu.ie

Abstract. The use and application of virtual reality (VR) continue to grow as more advanced VR-capable hardware is developed. With VR hardware entering the mainstream, it becomes increasingly important to develop methods to support use cases previously performed using conventional user interfaces, such as desktop systems with keyboard and mouse. One very important such use case is video retrieval and browsing. Previous experiments have indicated that, while there appears to be a benefit to browsing videos in VR, the expression of textual queries remains a challenge.
In this paper, we describe a new version of the vitrivr-VR virtual reality multimedia retrieval system, as it will participate in the video browser showdown (VBS) 2025. This new version of vitrivr-VR aims to overcome the two main challenges that have been identified in participation in previous installments of the VBS, namely the retrieval performance and the expression of textual queries, while continuing to leverage the advantages of video browsing in VR. We plan to overcome these challenges using a more performant retrieval backend (vitrivr-engine) and the use of newer VR-capable hardware (Apple Vision Pro) to enable easier text input.

Keywords: Video Browser Showdown · Virtual Reality · Interactive Video Retrieval · Content-based Retrieval.

1 Introduction

Virtual reality (VR) remains a rapidly developing technology. As a result of the investments and continued support of several large consumer electronics manufacturers, the boundaries of what VR hardware is capable of are pushed at a rapid pace. Currently, VR is most frequently used for entertainment purposes, such as video games and media consumption. However, for VR to be accepted as a user interface modality equivalent in utility to conventional desktop systems, methods need to be developed to enable use cases beyond these.

One such use case is video retrieval and browsing. With the large amount of video content recorded and uploaded to the internet or stored in private collections, methods to effectively search and browse such collections are required. A

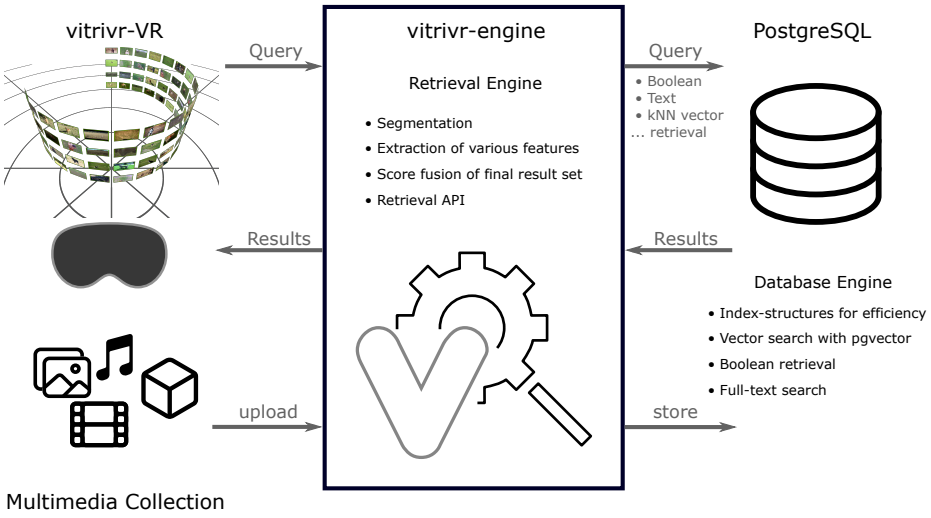

Fig. 1. System architecture of vitrivr-VR.

vibrant research community has emerged around developing such methods, leading to community-run evaluation campaigns, such as the Video Browser Showdown (VBS). While the majority of participants at the VBS have historically been desktop-based systems, VR systems are no longer uncommon. vitrivr-VR is one such VR multimedia retrieval system, having participated in several previous installments of the VBS [9]. Two main challenges have been identified in the context of vitrivr-VR at previous installments of the VBS: low retrieval performance for large datasets and slow, error-prone text input for textual queries [7]. With the expansion of both the V3C [6] to include all three shards and of the marine video kit dataset [11] used at the VBS 2025, high retrieval performance is of even greater relevance at this installment, and with the continued dominance of multi-modal visual-text embeddings for multimedia retrieval, fast, reliable text input continues to be of great importance.

In this paper, we describe a version of vitrivr-VR developed to overcome the two main challenges identified in previous evaluation campaigns. To overcome the challenge of retrieval performance in the face of increasing dataset sizes, we switch to vitrivr-engine [1], a new retrieval backend supporting much more efficient nearest neighbor search. By using the Apple Vision Pro headset, we are able to overcome the text input limitations of VR through improved, on-device speech-to-text, as well as the use of physical keyboards, thanks to the headset's mixed reality capabilities.

2 vitrivr-VR

vitrivr-VR is a prototype virtual reality multimedia retrieval and analytics system based on the vitrivr system stack [5]. It consists of a three-tier architecture

consisting of a database, feature extraction and retrieval engine, and virtual reality user interface. Each of these three parts of the vitrivr-VR architecture has been significantly updated or replaced in comparison to previous participations in the VBS. In this new version of vitrivr-VR, we use PostgreSQL as a database with the pgvector[3] extension for efficient nearest neighbor search [4]. As the interface to the PostgreSQL database and to facilitate feature extraction and retrieval logic, we use vitrivr-engine. The virtual reality user interface had been rebuilt specifically for the Apple Vision Pro, allowing it to run directly on the headset. A diagram of the system architecture of vitrivr-VR is presented in Figure 1.

The new architecture of vitrivr-VR allows us to address the two main challenges encountered by vitrivr-VR in previous evaluation campaigns. In the following sections, we highlight how the new components enable faster retrieval times, even for large datasets, as well as more efficient and effective text entry.

2.1 Retrieval Performance

While interactive retrieval in very large datasets depends on many other factors, such as the accessibility of the user interface and the available methods for browsing in query results, retrieval performance plays a very important role in a system's efficacy. Especially in cases where many iterations of queries are necessary, the delay with which query results are returned has a big impact, not only on the frequency with which an operator can iterate but possibly also on how often the operator may choose to reformulate a query. Such query reformulation is particularly important for textual known item search (T-KIS) tasks, such as those used in the VBS, which reveal additional information over the course of a task.

During previous participation in the VBS, retrieval performance for vitrivr-VR has been relatively poor. In several cases, other participating systems were able to submit correct results before the first set of results could be retrieved by vitrivr-VR [3,7]. Especially for very large datasets, like the full V3C [6] dataset that will be used at the VBS 2025, high retrieval performance is crucial to solve interactive retrieval tasks.

To address this challenge, we use vitrivr-engine in combination with PostgreSQL using the pgvector extension as our retrieval backend. The vitrivr-engine feature extraction and retrieval engine allows flexible and customizable specifications of feature extraction and retrieval pipelines. Through this flexibility, the specific retrieval needs of vitrivr-VR can be addressed in a much more specific and individual way. The main performance improvement comes from the use of PostgreSQL with the pgvector extension as a vector store and for nearest neighbor search. As pgvector allows parallelization and a variety of indexes for vector search queries, retrieval performance is high, even for very large datasets.

[3] https://github.com/pgvector/pgvector

2.2 Text Input

Text input in virtual reality remains a significant challenge [10]. An analysis of the performance of vitrivr-VR at the VBS 2023 indicates that the textual query formulation in VR takes at least 24% longer on average than using an equivalent desktop-based system [7]. As this number also includes the time taken to understand the competition task and for the query to be executed, which may be assumed to be similar for both systems, this difference is likely much larger. This has also been supported by qualitative data collected at previous evaluation campaigns.

In a previous approach to tackling this challenge, we attempted to overcome the slow text input speed in VR by splitting the system into two components: a desktop query formulation interface and a VR multimedia browsing interface [8]. While this approach greatly improved text input speed, the additional communication overhead between the two operators likely led to the observed overall decreased performance.

In this new version of vitrivr-VR, we try to combine these previous approaches by utilizing the Apple Vision Pro's mixed reality capabilities. By developing vitrivr-VR directly for the Apple Vision Pro, we are able to utilize a physical keyboard for textual query formulation while preserving the benefits of video browsing in virtual reality.

This version of vitrivr-VR will primarily use textual queries based primarily on OCR, ASR, and OpenCLIP [2]. We use OCR and ASR results to filter the collection and perform a nearest-neighbor search using OpenCLIP on the resulting video segments.

3 Conclusion

In this paper, we introduce the version of vitrivr-VR with which we intend to participate in the VBS 2025. We address the two issues most commonly identified for vitrivr-VR in previous participations in evaluation campaigns: retrieval performance and text input. We address these with a new retrieval backend consisting of vitrivr-engine and PostgreSQL, as well as developing the user interface natively for the Apple Vision Pro to allow text input using a physical keyboard.

Acknowledgments. This work was partly supported by the Swiss National Science Foundation through projects "Participatory Knowledge Practices in Analog and Digital Image Archives" (contract no. 193788) and MediaGraph (contract no. 202125), and the Horizon Europe project XReco, funded by a grant from the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00268.

References

1. Gasser, R., Arnold, R., Faber, F., Schuldt, H., Waltenspül, R., Rossetto, L.: A new retrieval engine for vitrivr. In: Rudinac, S., Hanjalic, A., Liem, C.C.S., Worring, M.,

Jónsson, B.Þ., Liu, B., Yamakata, Y. (eds.) MultiMedia Modeling - 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 - February 2, 2024, Proceedings, Part IV. Lecture Notes in Computer Science, vol. 14557, pp. 324–331. Springer (2024). https://doi.org/10.1007/978-3-031-53302-0_28
2. Ilharco, G., Wortsman, M., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: OpenCLIP. Zenodo (2021). https://doi.org/10.5281/ZENODO.5143773
3. Lokoč, J., Andreadis, S., Bailer, W., Duane, A., Gurrin, C., Ma, Z., Messina, N., Nguyen, T.N., Peška, L., Rossetto, L., Sauter, L., Schall, K., Schoeffmann, K., Khan, O.S., Spiess, F., Vadicamo, L., Vrochidis, S.: Interactive video retrieval in the age of effective joint embedding deep models: Lessons from the 11th VBS. Multimedia Systems **29**(6), 3481–3504 (2023). https://doi.org/10.1007/s00530-023-01143-5
4. Rossetto, L., Gasser, R.: Feature-driven video segmentation and advanced querying with vitrivr-engine. In: MultiMedia Modeling - 31st International Conference, MMM 2024, Nara, Japan, January 7 - 10, 2024. Lecture Notes in Computer Science, Springer (2025)
5. 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. pp. 1183–1186. ACM, Amsterdam The Netherlands (2016). https://doi.org/10.1145/2964284.2973797
6. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C – A Research Video Collection. In: MultiMedia Modeling, vol. 11295, pp. 349–360. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-05710-7_29
7. Spiess, F., Gasser, R., Heller, S., Schuldt, H., Rossetto, L.: A Comparison of Video Browsing Performance between Desktop and Virtual Reality Interfaces. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. pp. 535–539. ACM, Thessaloniki Greece (2023). https://doi.org/10.1145/3591106.3592292
8. Spiess, F., Gasser, R., Schuldt, H., Rossetto, L.: The Best of Both Worlds: Lifelog Retrieval with a Desktop-Virtual Reality Hybrid System. In: Proceedings of the 6th Annual ACM Lifelog Search Challenge. pp. 65–68. ACM, Thessaloniki Greece (2023). https://doi.org/10.1145/3592573.3593107
9. Spiess, F., Rossetto, L., Schuldt, H.: Exploring Multimedia Vector Spaces with vitrivr-VR. In: MultiMedia Modeling, vol. 14557, pp. 317–323. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-53302-0_27
10. Spiess, F., Weber, P., Schuldt, H.: Direct Interaction Word-Gesture Text Input in Virtual Reality. In: 2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR). pp. 140–143. IEEE, CA, USA (2022). https://doi.org/10.1109/AIVR56993.2022.00028
11. Truong, Q., Vu, T., Ha, T., Lokoc, J., Tim, Y.H.W., Joneja, A., Yeung, S.: Marine video kit: A new marine video dataset for content-based analysis and retrieval. In: Dang-Nguyen, D., Gurrin, C., Larson, M.A., Smeaton, A.F., Rudinac, S., Dao, M., Trattner, C., Chen, P. (eds.) MultiMedia Modeling - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023, Proceedings, Part I. Lecture Notes in Computer Science, vol. 13833, pp. 539–550. Springer (2023). https://doi.org/10.1007/978-3-031-27077-2_42