Towards Explainable Interactive Multi-Modal Video Retrieval with vitrivr

Silvan Heller[0000−0001−5386−330X], Ralph Gasser[0000−0002−3016−1396],
Cristina Illi[0000−0001−9217−4662], Maurizio Pasquinelli[0000−0003−2076−8327],
Loris Sauter[0000−0001−8046−0362], Florian Spiess[0000−0002−3396−1516], and
Heiko Schuldt[0000−0001−9865−6371]

Department of Mathematics and Computer Science
University of Basel, Basel, Switzerland
{firstname.lastname}@unibas.ch

Abstract. This paper presents the most recent iteration of the vitrivr multimedia retrieval system for its participation in the Video Browser Showdown (VBS) 2021. Building on existing functionality for interactive multi-modal retrieval, we overhaul query formulation and results presentation for queries which specify temporal context, extend our database with index structures for similarity search and present experimental functionality aimed at improving the explainability of results with the objective of better supporting users in the selection of results and the provision of relevance feedback.

Keywords: Video Browser Showdown · Interactive Video Retrieval · Content-based Retrieval · Explainability

1 Introduction

The Video Browser Showdown (VBS) [17] is a major evaluation campaign for *interactive video retrieval* and celebrates its 10^{th} anniversary in 2021. The vitrivr system and its predecessors have been long-running participants to VBS [19], utilising different methods from sketch-based queries to deep-learning methods for concept detection, and text retrieval for OCR and ASR in recent years [20,22].

vitrivr always had a focus on multi-modal multimedia retrieval, enabling users to mix and match different modalities to find a particular item in a collection. Being a general-purpose retrieval system, vitrivr has found many applications ranging, for example lifelog search [6,16].

In this paper, we present the vitrivr system as envisioned to participate at VBS'21, including its changes to temporal queries, newly added index structures and improvements towards the explainability of results. We have used the time since the last VBS to make improvements to query formulation and our retrieval model [8], to take another look at relevance feedback and results presentation [7], and to enhance our open source database, which is described in more detail in [3]. The backend and database layer are also used by another system participating at VBS, vitrivr-VR [23].

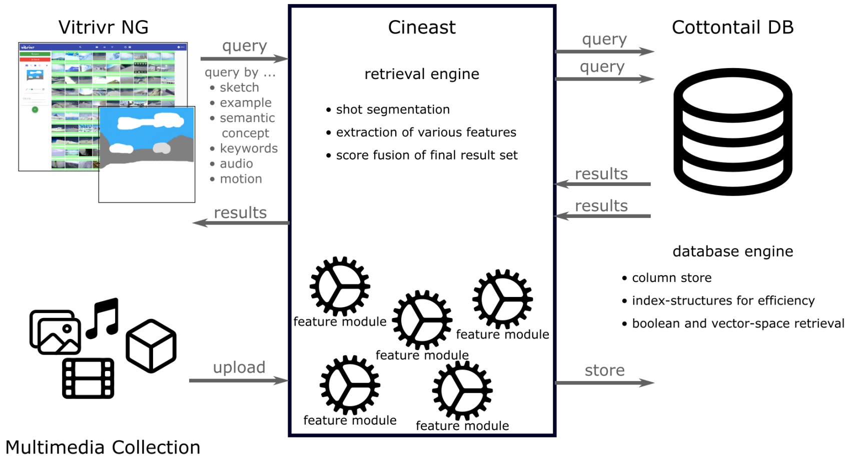

Fig. 1: Architecture overview of vitrivr [6]

The remainder of this paper is structured as follows: Section 2 gives an overview of the vitrivr system, Section 3 discusses our new approach to temporal scoring, Section 4 introduces the index structures we will be using for similarity search, Section 5 gives examples of the envisioned user-facing explanations of results and Section 6 concludes.

2 vitrivr

vitrivr[1] is an open source, content-based multimedia retrieval stack with explicit support for different modalities, such as images, audio, video, and 3D models. As such, vitrivr serves as a platform for a wide range of multimedia research activities such as gesture retrieval, search and exploration in VR and cultural heritage applications. The stack covers all aspects of multimedia retrieval, namely feature extraction and storage, content serving, query formulation and execution, and result presentation. In Figure 1, we provide an architectural overview of vitrivr and its three main components: The storage layer Cottontail DB [3], the retrieval engine Cineast [18] and the presentation layer vitrivr-ng.

In vitrivr, many different kinds of features and data sources are used. We extract a variety of low-level visual features for color and texture, use more traditional features such as SURF [1], deep features [9] and textual information (ASR, OCR) for fulltext search [20]. For an in-depth discussion of the retrieval model unifying different types of features, we refer to [8].

While vitrivr has always been focused on *retrieval*, we have recently experimented with different presentation modes and relevance feedback for a stronger

[1] https://vitrivr.org/

focus on *exploration*, which has shown promising results for Ad-Hoc Video Search (AVS)-style tasks [7].

3 Temporal Querying

As discussed in [6,8], users of vitrivr can currently formulate temporal queries by combining multiple, independent queries. Other systems at VBS employ different approaches to the same problem [13,14]. In addition, users must specify a temporal bound wherein potential matches should occur, e.g., results of query A within 30 seconds of query B. The two queries are then processed independently in parallel by the retrieval engine. The results are combined by the front-end such that the temporal proximity of matching pairs in both result sets within the specified limits lead to a higher overall score.

Evaluations with different versions of vitrivr [7] and against other systems have shown that user experience and retrieval speed deteriorates quickly for large result sets. Furthermore, user studies have found that users almost never submit more than two queries jointly, even if they see longer portions of a video and the general feedback indicates that users find the current presentation as well as query formulation to be very unintuitive.

This feedback, combined with the proven effectiveness of temporal querying, has led to revisiting this functionality. Aggregating results by temporal closeness in the front-end suffers from inherent architectural limitations: Since Cineast only sends a limited number of results per query, some segments might be present in multiple result lists, but get lost due to the cutoff. For this iteration of Cineast, we therefore plan to compute matches for temporal queries already in the retrieval engine to address this issue and improve performance.

Additionally, we are experimenting with simpler ways to enable users to express temporal dependencies while preserving the multi-modality and expressiveness of the current front-end, for instance by using different ways of presenting matches for temporal queries.

4 Index Structures for Similarity Search

Already with the first shard of the V3C [21] dataset – V3C1 [2] – fitting all relevant data into main memory on commodity hardware is challenging. vitrivr uses Cottontail DB [3] as its database, which makes it well-positioned for future increase in dataset size. However, as these datasets become larger, linear nearest neighbor search (NNS) quickly becomes a bottleneck, especially in a time-critical and competitive setting such as VBS, where every second counts.

Promising techniques for approximate NNS that we currently consider are LSH-based methods [10], PQ [11] and the more recent ScANN [4] algorithm. Due to the curse of dimensionality, however, accelerating NNS with such indexes often comes at the cost of sacrificing either execution performance or accuracy. We will therefore focus on ways to optimize the trade-off between the two, e.g.,

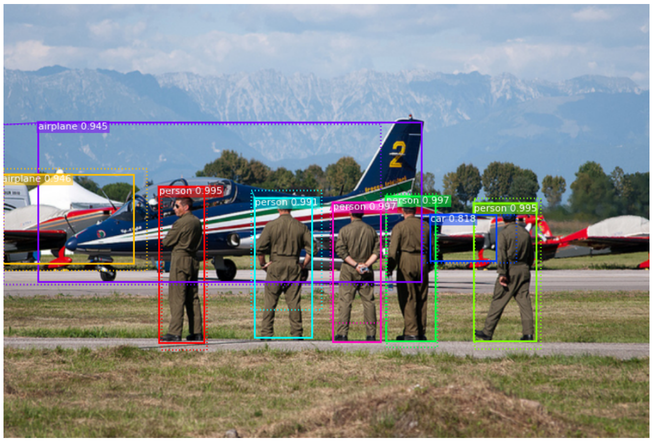

Fig. 2: Concept bounding boxes can be used to inspire the user when improving the query [5].

by combining different index structures or by progressively improving result sets, as more accurate information becomes available after some time.

5 Towards the Explainability of Search Results

The topic of explainability has become increasingly important in recent years for two reasons: First, understanding the inner workings of machine learning algorithms, which are often considered to be black boxes, helps developers to improve their systems. Second, and often overlooked, by providing information on the retrieval process, users can improve their queries in an interactive and iterative way, which is exactly the goal of interactive retrieval. In contrast to relevance feedback which aims at improving results based on feedback by the user (as done in i.e. [12]), the objective of explaining results is to help users improve their query based on the information conveyed by the result.

Displaying explanations for different modalities from those employed by the user during their search can also inspire them to think of new ways of modifying their query. Especially so, if the user is stuck and does not get satisfying results with the current modality. These visualizations can then be used interactively to improve query formulation, for example, by performing more-like-this searches on specific objects or proposing concepts users had previously not considered to the query. As an example, consider fig. 2, where detected concepts are highlighted with their bounding boxes [5,15].

vitrivr aims to explain results in two ways: First, on the level of an individual result by showing precisely why a feature was considered a result a match, e.g., by showing feature visualizations And second, on a result set level by giving users information about properties of all returned results such as co-occurring concepts and frequently occurring textual descriptions.

6 Conclusion

In this paper, we presented the current iteration of the vitrivr system participating at VBS 2021. Compared to previous participations, vitrivr will add improvements for temporal queries, provide additional index structures at database level, and improve the overall query process and enhance the user experience by explaining the retrieval results.

Acknowledgements

This work was partly supported by the Hasler Foundation in the context of the project City-Stories (contract no. 17055) and by the Swiss National Science Foundation, project Polypheny-DB (contract no. 200021_172763).

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding **110**(3) (2008)
2. Berns, F., Rossetto, L., Schoeffmann, K., Beecks, C., Awad, G.: V3c1 dataset: An evaluation of content characteristics. In: Proceedings of the International Conference on Multimedia Retrieval (2019)
3. 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)
4. Guo, R., Sun, P., Lindgren, E., Geng, Q., Simcha, D., Chern, F., Kumar, S.: Accelerating large-scale inference with anisotropic vector quantization. arXiv preprint arXiv:1908.10396 (2020)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE international conference on computer vision (2017)
6. 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, LSC@ICMR 2020, Dublin, Ireland (2020)
7. 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 (SISAP) (2020)
8. Heller, S., Sauter, L., Schuldt, H., Rossetto, L.: Multi-stage queries and temporal scoring in vitrivr. In: IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (2020)
9. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
10. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing (1998)
11. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(1) (2010)

12. Jónsson, B.Þ., Khan, O.S., Ragnarsdóttir, H., Þorleiksdóttir, Þ., Zahálka, J., Rudinac, S., Guðmundsson, G.Þ., Amsaleg, L., Worring, M.: Exquisitor: Interactive learning at large. arXiv preprint arXiv:1904.08689 (2019)
13. Lokoc, J., Sourcek, T., Vesely, P., Mejzlik, F., Ji, J., Xu, C., Li, X.: A w2vv++ case study with automated and interactive text-to-video retrieval. In: Proceedings of the 28 ACM International Conference on Multimedia (2020)
14. Nguyen, P.A., Wu, J., Ngo, C., Francis, D., Huet, B.: VIREO @ video browser showdown 2020. In: International Conference on Multimedia Modeling MMM (2020)
15. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. CoRR **abs/1506.01497** (2015), http://arxiv.org/abs/1506.01497
16. Rossetto, L., Gasser, R., Heller, S., Amiri Parian, M., Schuldt, H.: Retrieval of structured and unstructured data with vitrivr. In: Proceedings of the ACM Workshop on Lifelog Search Challenge (2019)
17. 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)
18. 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. pp. 336–341. Springer (2016)
19. Rossetto, L., Giangreco, I., Heller, S., Tănase, C., Schuldt, H., Dupont, S., Seddati, O., Sezgin, M., Altıok, O.C., Sahillioğlu, Y.: Imotion–searching for video sequences using multi-shot sketch queries. In: International Conference on Multimedia Modeling MMM (2016)
20. 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 MMM (2019)
21. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C–A Research Video Collection. In: International Conference on Multimedia Modeling MMM (2019)
22. 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 MMM (2020)
23. Spiess, F., Gasser, R., Heller, S., Rossetto, L., Sauter, L., Schuldt, H.: Competitive interactive video retrieval in virtual reality with vitrivr-vr. In: International Conference on Multimedia Modeling MMM (2021)