Exploring Effective Interactive Text-based Video Search in vitrivr

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

[1] University of Basel, Basel, Switzerland
{firstname}.{lastname}@unibas.ch
[2] University of Zurich, Zurich, Switzerland
rossetto@ifi.uzh.ch

Abstract. vitrivr is a general purpose retrieval system that supports a wide range of query modalities. In this paper, we briefly introduce the system and describe the changes and adjustments made for the 2023 iteration of the video browser showdown. These focus primarily on text-based retrieval schemes and corresponding user-feedback mechanisms.

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

1 Introduction

The Video Browser Showdown (VBS) [9,14,19] is a long-running evaluation campaign for interactive multimedia retrieval and user-centric video search. Since 2012 [27], the VBS has provided a highly competitive setup in which systems and their operators are tasked to find video segments within a large collection. The collection currently used is a subset of the Vimeo Creative Commons Collection (V3C) [26], which comprises 2300 h of video material that totals to 1.6 TB in size. In addition to V3C1 [3] and V3C2 [25], the 2023 installment of VBS will feature a homogeneous underwater / scuba diving dataset called the Marine Video Kit [30], of roughly 230 GB and a duration of approximately 11.5 h.

The VBS consists of two types of tasks: Known-Item Search (KIS) and Ad-hoc Video Search (AVS) [12]. The former involves finding a specific video segment based on either a visual preview or a textual description. The latter requires finding items of interest that match a more general description (their correctness is manually judged during the competition).

In this paper, we present vitrivr – an open-source content-based multi-modal multimedia retrieval system – and improvements made to it compared to previous iterations. The 2023 installment marks the 9$^{\text{th}}$ time[3] vitrivr participates to VBS in a row [7], with two winning participations in the last four years [8,24].

[3] including its predecessor, the iMotion system

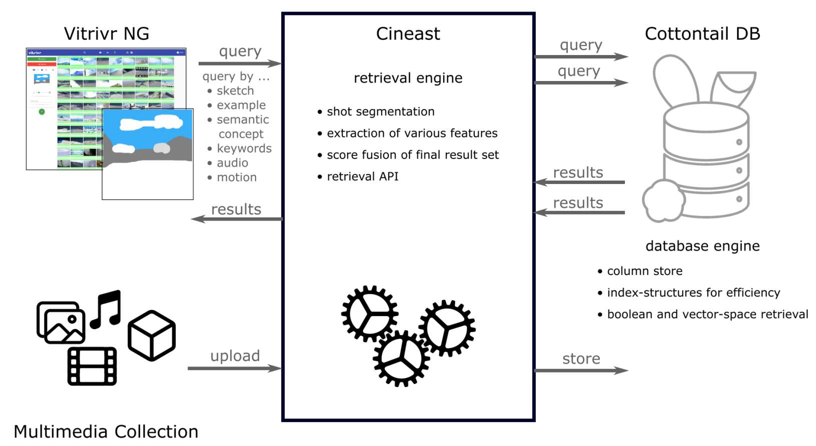

Fig. 1. System overview for vitrivr and its three major components: vitrivr-ng, Cineast and Cottontail DB. Slightly modified version of [8, Fig. 1].

The remainder of this paper is structured as follows: Section 2 provides an overview of the vitrivr stack, Section 3 highlights the various additions to the stack and we conclude the paper in Section 4.

2 vitrivr

vitrivr [23] is an open-source multimedia retrieval stack, capable of supporting a broad range of media and query types including, but not limited to, video (search). An overview of vitrivr's architecture is provided in Figure 1 [8, Fig 1]. The stack is composed of three main components:

Cottontail DB [5] is the database layer of vitrivr and can be used to store, manage, and query scalar metadata as well as high-dimensional feature vectors. Cottontail DB allows for efficient similarity and Boolean retrieval.

Cineast [17,21,22] is the feature extraction and retrieval engine of the stack. It generates the different feature representation from the input data (videos, user provided queries), orchestrates query execution and implements result aggregation and score fusion.

vitrivr-ng [6] is vitrivr's web-based user interface and facilitates query formulation, result presentation and allows for efficient exploration. In addition, it also enables late-stage filtering and fusion.

All of vitrivr's components are freely available from the project website.[4] Some of the aforenamed components also serve as the basis for other multimedia retrieval systems such as vitrivr-vr [28,29] and Lifegraph [18].

[4] https://vitrivr.org

3 Novelties for VBS 2023

Since the previous iteration of VBS has shown [9] that most of the top performing systems rely on video-text co-embeddings and support some form of temporal query, we have put some focus on refining and extending means that allow for text-based search.

3.1 Improved Visual-Text Co-Embedding

We introduced the first version of our visual-text co-embedding in [29], which consisted of a shallow network that projects the output of two uni-modal pre-trained backbones into a common, semantically aligned space. In order to handle multiple frames of a video rather than only a single image, all frames are passed through their visual backbone and its output is pooled before projection. For this year's iteration of VBS, we have updated several aspects of this approach. The two backbones have been replaced and the aggregation scheme and projection methods have been refined. For textual embedding, we now use a multilingual backbone [31] in order to increase accessibility to non-native English speakers. For the visual frame-level embeddings, we use a more recent convolutional architecture [11] and remove both the final classification as well as the spatial pooling layers. Inspired by [2], we extend the aggregation scheme to not pool the visual embeddings indiscriminately but rather pay attention to the spatial and temporal origin of image-patch embeddings.

3.2 CLIP and vitrivr

The release of the CLIP [16] model by OpenAI in 2021 marked a step-change in the quality achievable when searching for images using text describing their semantic content. In the 2022 edition of VBS, several of the highest scoring teams relied at least in part on feature representations generated by CLIP [1,10,13]. In view of this decisive demonstration of its effectiveness, we have added a CLIP-based feature extractor to vitrivr for the 2023 edition of VBS. The features are extracted based on only one representative frame per shot, as provided by the dataset [26]. During runtime, we provide users the means to chose our co-embeddings or CLIP as query handler.

Since prompt-engineering appears to be a relevant factor in the effective performance of contemporary joint visual language models, as minor changes in the input can lead to rather substantial changes in returned results in some cases, we also employ CLIP-guided image captioning methods, specifically [4] and [15], in order to generate one caption (each) per representative frame of every shot. These captions are not intended to be used for search directly (although such functionality is also supported) but rather to provide feedback to an operator, what a reasonable textual query would be to retrieve any given result. This feedback mechanism can be used by operators to familiarize themselves with the intricacies of the feature in order to help them to construct more effective search prompts.

3.3 SIMD support for Cottontail DB

Query execution speed is of the essence for interactive video retrieval, especially in competitive settings such as VBS. In the latest iteration of Cottontail DB [5] — the multimedia database layer used by vitrivr — we have therefore started to exploit the use of SIMD instructions to accelerate query execution for brute-force search. The explicit use of SIMD has been enabled by the recent incubation of the new Java Vector API proposed in the JEPs 338, 414 and 417.[5] Even though the current implementation is rather straightforward and despite the feature still being in an early beta stadium, we can report a speed-up of between 20-30% especially for high-dimensional vectors ($d > 1024$). We expect to attain even more acceleration by transitioning the underlying query execution engine from an iterator to a batched processing-model in the (not too distant) future.

3.4 Human-in-the-loop

A vital component in user-centric video search is the human (retrieval) system operator. The vitrivr team employs regular VBS-style dry-runs since quite some time with a dedicated evaluation setup. In this year's installment, we use our own deployment of the DRES [20] system, with tasks specifically created for that purpose. We analyse each dry-run and, particularly analyse those tasks, where we — the system operator and system — could not find the target. As outlined in Section 3.2 we use, among others, the means of textual feature representation of the target in order to learn what search terms would have been purposeful. Specifically for the 2023 installment, we also will have dry-runs in December with peers who did not work on vitrivr, to simulate novice sessions and we might need to adjust some of the user interface functionality based on their feedback.

4 Conclusion

In this paper, we presented the version of vitrivr with which we plan to participate at VBS 2023. As has recent analysis shown, the current trend in user-centric video search goes towards deep learning supported video-text co-embeddings such as CLIP. Thus, we focus on improvements in this domain by expanding our visual-text co-embedding — among others — with a multilingual backbone and introduce the CLIP model in our pipeline as well. Furthermore, various parts of the open-source retrieval system vitrivr have been improved and we will systematically train our system operators with the system as well as simulate novice sessions with our local peers.

Acknowledgements

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).

[5] See https://openjdk.org/jeps/338, accessed September 2022.

References

1. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: VISIONE at video browser showdown 2022. In: MultiMedia Modeling. Springer (2022). https://doi.org/10.1007/978-3-030-98355-0_52
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: International Conference on Computer Vision, ICCV (2021). https://doi.org/10.1109/ICCV48922.2021.00175
3. Berns, F., Rossetto, L., Schoeffmann, K., Beecks, C., Awad, G.: V3C1 Dataset: An Evaluation of Content Characteristics. In: International Conference on Multimedia Retrieval. ACM (2019). https://doi.org/10.1145/3323873.3325051
4. Cho, J., Yoon, S., Kale, A., Dernoncourt, F., Bui, T., Bansal, M.: Fine-grained image captioning with CLIP reward. In: Findings of the Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.findings-naacl.39
5. Gasser, R., Rossetto, L., Heller, S., Schuldt, H.: Cottontail DB: An Open Source Database System for Multimedia Retrieval and Analysis. In: International Conference on Multimedia. ACM (2020). https://doi.org/10.1145/3394171.3414538
6. Gasser, R., Rossetto, L., Schuldt, H.: Multimodal Multimedia Retrieval with vitrivr. In: International Conference on Multimedia Retrieval (2019)
7. Heller, S., Arnold, R., Gasser, R., Gsteiger, V., Parian-Scherb, M., Rossetto, L., Sauter, L., Spiess, F., Schuldt, H.: Multi-modal Interactive Video Retrieval with Temporal Queries. In: MultiMedia Modeling. Springer (2022). https://doi.org/10.1007/978-3-030-98355-0_44
8. Heller, S., Gasser, R., Illi, C., Pasquinelli, M., Sauter, L., Spiess, F., Schuldt, H.: Towards Explainable Interactive Multi-modal Video Retrieval with Vitrivr. In: MultiMedia Modeling. Springer (2021). https://doi.org/10.1007/978-3-030-67835-7_41
9. Heller, S., Gsteiger, V., Bailer, W., Gurrin, C., Jónsson, B.Þ., Lokoč, J., Leibetseder, A., Mejzlík, F., Peška, L., Rossetto, L., Schall, K., Schoeffmann, K., Schuldt, H., Spiess, F., Tran, L.D., Vadicamo, L., Veselý, P., Vrochidis, S., Wu, J.: Interactive video retrieval evaluation at a distance: Comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown. International Journal of Multimedia Information Retrieval **11**(1), 1–18 (2022). https://doi.org/10.1007/s13735-021-00225-2
10. Hezel, N., Schall, K., Jung, K., Barthel, K.U.: Efficient search and browsing of large-scale video collections with vibro. In: MultiMedia Modeling. Springer (2022). https://doi.org/10.1007/978-3-030-98355-0_43
11. Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. CoRR **abs/2201.03545** (2022)
12. Lokoč, J., Bailer, W., Barthel, K.U., Gurrin, C., Heller, S., þór Jónsson, B., Peška, L., Rossetto, L., Schoeffmann, K., Vadicamo, L., Vrochidis, S., Wu, J.: A Task Category Space for User-Centric Comparative Multimedia Search Evaluations. In: MultiMedia Modeling (2022). https://doi.org/10.1007/978-3-030-98358-1_16
13. Lokoc, J., Mejzlík, F., Soucek, T., Dokoupil, P., Peska, L.: Video search with context-aware ranker and relevance feedback. In: MultiMedia Modeling. Springer (2022). https://doi.org/10.1007/978-3-030-98355-0_46
14. Lokoč, J., Veselý, P., Mejzlík, F., Kovalčík, G., Souček, T., Rossetto, L., Schoeffmann, K., Bailer, W., Gurrin, C., Sauter, L., Song, J., Vrochidis, S., Wu, J., þóR Jónsson, B.: Is the Reign of Interactive Search Eternal? Findings from the Video Browser Showdown 2020. ACM Transactions on Multimedia Computing, Communications, and Applications (2021). https://doi.org/10.1145/3445031

15. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: CLIP prefix for image captioning. CoRR **abs/2111.09734** (2021)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
17. Rossetto, L.: Multi-Modal Video Retrieval. Ph.D. thesis, University of Basel (2018)
18. Rossetto, L., Baumgartner, M., Gasser, R., Heitz, L., Wang, R., Bernstein, A.: Exploring Graph-querying approaches in LifeGraph. In: Workshop on Lifelog Search Challenge (2021). https://doi.org/10.1145/3463948.3469068
19. Rossetto, L., Gasser, R., Lokoč, J., Bailer, W., Schoeffmann, K., Muenzer, B., Souček, T., Nguyen, P.A., Bolettieri, P., Leibetseder, A., Vrochidis, S.: Interactive Video Retrieval in the Age of Deep Learning – Detailed Evaluation of VBS 2019. IEEE Transactions on Multimedia (2021)
20. Rossetto, L., Gasser, R., Sauter, L., Bernstein, A., Schuldt, H.: A system for interactive multimedia retrieval evaluations. In: MultiMedia Modeling - 27th International Conference, MMM 2021, Prague, Czech Republic, June 22-24, 2021, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12573, pp. 385–390. Springer (2021)
21. Rossetto, L., Giangreco, I., Heller, S., Tanase, C., Schuldt, H.: Searching in Video Collections Using Sketches and Sample Images - The Cineast System. In: MultiMedia Modeling. Springer (2016). https://doi.org/10.1007/978-3-319-27674-8_30
22. Rossetto, L., Giangreco, I., Schuldt, H.: Cineast: A multi-feature sketch-based video retrieval engine. In: International Symposium on Multimedia (2014)
23. Rossetto, L., Giangreco, I., Tanase, C., Schuldt, H.: vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections. In: ACM Conference on Multimedia (2016). https://doi.org/10.1145/2964284.2973797
24. Rossetto, L., Parian, M.A., Gasser, R., Giangreco, I., Heller, S., Schuldt, H.: Deep Learning-Based Concept Detection in vitrivr. In: MultiMedia Modeling. Springer (2019). https://doi.org/10.1007/978-3-030-05716-9_55
25. Rossetto, L., Schoeffmann, K., Bernstein, A.: Insights on the V3C2 Dataset. CoRR **abs/2105.01475** (2021)
26. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C - A research video collection. In: MultiMedia Modeling. Springer (2019). https://doi.org/10.1007/978-3-030-05710-7_29
27. Schoeffmann, K.: Video Browser Showdown 2012-2019: A Review. In: International Conference on Content-Based Multimedia Indexing (2019)
28. Spiess, F., Gasser, R., Heller, S., Parian-Scherb, M., Rossetto, L., Sauter, L., Schuldt, H.: Multi-modal Video Retrieval in Virtual Reality with vitrivr-VR. In: MultiMedia Modeling (2022). https://doi.org/10.1007/978-3-030-98355-0_45
29. Spiess, F., Gasser, R., Heller, S., Rossetto, L., Sauter, L., Schuldt, H.: Competitive Interactive Video Retrieval in Virtual Reality with vitrivr-VR. In: MultiMedia Modeling. Springer (2021). https://doi.org/10.1007/978-3-030-67835-7_42
30. Truong, Q.T., Vu, T.A., Ha, T.S., Lokoc, J., Tim, Y.H.W., Joneja, A., Yeung, S.K.: Marine video kit: A new marine video dataset for content-based analysis and retrieval. In: MultiMedia Modeling - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023. Lecture Notes in Computer Science, Springer (2023)
31. Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G.H., Yuan, S., Tar, C., Sung, Y.H., Strope, B., Kurzweil, R.: Multilingual universal sentence encoder for semantic retrieval (2019). https://doi.org/10.48550/ARXIV.1907.04307