Multi-Modal 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]], Mahnaz Parian-Scherb[1 [0000−0001−7063−8585]], 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. In multimedia search, appropriate user interfaces (UIs) are essential to enable effective specification of the user's information needs and the user-friendly presentation of search results. vitrivr-VR addresses these challenges and provides a novel Virtual Reality-based UI on top of the multimedia retrieval system vitrivr. In this paper we present the version of vitrivr-VR participating in the Video Browser Showdown (VBS) 2022. We describe our visual-text co-embedding feature and new query interfaces, namely text entry, pose queries and temporal queries.

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

1 Introduction

User interaction plays an important role in multimedia search, although it is often underestimated. This concerns, on the one hand, the as precise and natural as possible specification of the information need for the search and, on the other hand, the presentation of, and ideally also the interaction with the search results. Virtual Reality (VR) based user interfaces (UIs) allow for innovative solutions to both phases: queries can be specified more naturally than in desktop UIs, and the VR-based results display goes way beyond a traditional 2D layout and allows users to immersively explore their search results.

These advantages are explored by vitrivr-VR, which combines a VR-based user interface with the vitrivr stack[3], an open-source multimedia retrieval system. One way to evaluate the effectiveness of novel retrieval systems is through evaluation campaigns like the Video Browser Showdown (VBS) [10], in which other VR systems, such as EOLAS [16], also participate.

[3] https://www.vitrivr.org/

This paper describes the version of vitrivr-VR with which we plan to participate in the VBS 2022. We focus on the changes that have resulted directly from lessons learned during the previous participation of vitrivr-VR in the VBS 2021 [14], which was vitrivr-VR's first appearance. We expect our improvements to help cope with the large VBS'22 dataset, the premiere usage of the combined first two shards (V3C1 and V3C2) of the V3C dataset [12].

For query specification, we have improved the speech-to-text transformation for text queries. This is backed by a more advanced visual-text co-embedding compared to the VBS'21 version of vitrivr-VR. We also add support for temporal queries to the vitrivr-VR query interface, i.e., users can specify several subqueries with temporal dependencies. Finally, vitrivr-VR also supports pose queries, to allow querying for specific body poses.

The paper is structured as follows: Section 2 introduces vitrivr-VR and Section 3 describes our visual-text co-embedding feature. Section 4 describes new interfaces for query formulation in VR, including text entry, temporal queries, and pose queries, and Section 5 concludes.

2 vitrivr-VR

vitrivr-VR is an experimental, virtual reality multimedia retrieval system prototype based on the open-source vitrivr stack. The full stack of vitrivr-VR consists of three main parts:

Cottontail DB [4] is a column store used in the vitrivr stack to store extracted multimedia features and metadata and perform Boolean as well as similarity queries. Cineast [11] is the retrieval engine and feature extractor at the core of the vitrivr stack. It facilitates both query processing and feature transformation during online retrieval as well as offline feature extraction phases. Additionally, Cineast is responsible for score fusion, result set reconstruction, and temporal scoring. The retrieval model is described further in [6]. vitrivr-VR is the experimental, VR-based user interface facilitating fully-immersive query formulation and results exploration in virtual reality, the newest iteration of which we describe in this paper. Within the vitrivr stack, it is an alternative to the conventional, web-based two-dimensional user interface vitrivr-ng.

vitrivr-VR is developed in Unity[4] with C# for the HTC Vive Pro and Valve Index, and communicates with Cineast using the RESTful API provided through its OpenAPI specifications. To interface with VR hardware, vitrivr-VR uses the Unity OpenXR plugin[5].

The user interface components of vitrivr-VR can be categorized into three main categories: i.) query formulation (see more details in Section 4), ii.) result set exploration, and iii.) media item inspection. The query modalities currently supported through query formulation are concept, text, Boolean, and geo-spatial queries.

[4] https://unity.com/
[5] https://docs.unity3d.com/Packages/com.unity.xr.openxr@latest

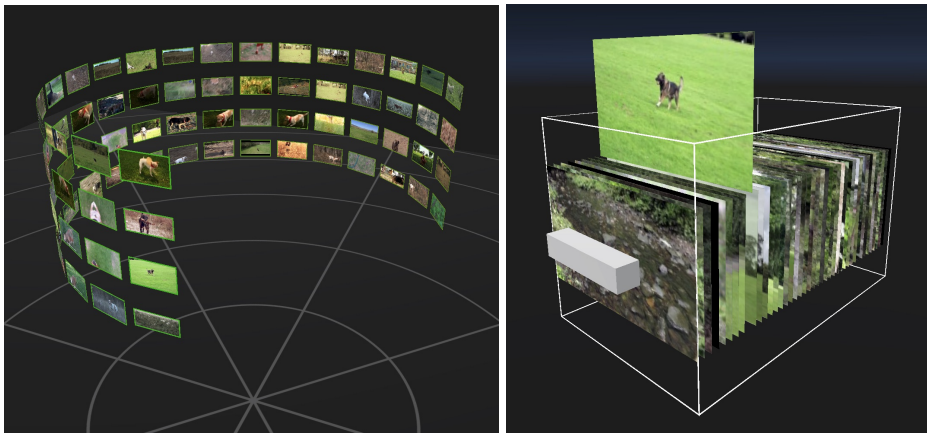

Fig. 1. Screenshots of the vitrivr-VR system; cylindrical results view (left) and video segment view (right).

Result set exploration is facilitated through a cylindrical, rotatable results display as seen in Figure 1, which supports both a media segment and a media object centered view. In the media segment centered view, each media segment result (shots in the case of videos) is displayed ordered by similarity score in a grid around the cylindrical display. The media object centered view displays only a configurable number of the top scoring segments for each media object and presents them with increasing distance from the cylindrical display with decreasing score. Segments from each media object occupy only a single grid position in the cylindrical display, ordered by score.

Media item inspection is done through the *media segment inspector* and the *media object segment view*. The media segment inspector pops out of the result view when a result is selected, can be freely moved and placed around the virtual space and persists even when the result set is cleared. It allows the selected media segment to be inspected in detail, i.e. in the case of video viewing the source video segment, as well as metadata such as detected concepts. The media object segment view is a temporally ordered, interactive 3D representation of a media object's segments, which can be accessed from the media segment inspector. It displays the segments inside a box, allowing users to quickly 'riffle' through them and select a segment for closer inspection in the media segment inspector. To increase precision for media objects containing large numbers of segments, a handle allows users to scale the box similar to pulling out a drawer.

3 Visual-Text Co-Embedding

The visual-text co-embedding feature, inspired by approaches such as [3,7], works as shown in Figure 2. For each video segment, each video frame is first transformed individually by a pre-trained visual feature encoder, and then aggre-

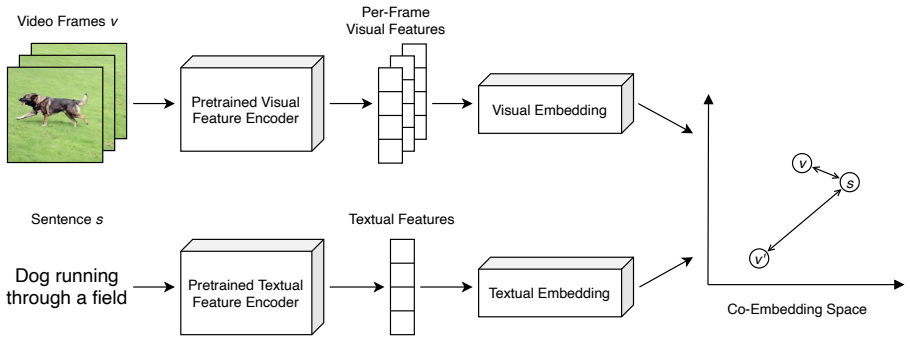

Fig. 2. The visual-text co-embedding as implemented in vitrivr. The objective is to minimize the distance between a video segment v and a matching query sentence s in the co-embedding space, while maximizing the distance to unrelated videos v'.

gated and embedded into a joint co-embedding space by an embedding network trained by us. The textual embedding path works analogously except no pooling is needed. For VBS'21, mean pooling was used for visual feature aggregation. We have improved the visual feature aggregation for VBS'22 to take the temporal context of video frames into account.

As a visual feature encoder, we use the output of the second to last layer of InceptionResNetV2 [15] trained on ImageNet [2] using average pooling through the Keras implementation[6]. As a textual feature encoder we use the Universal Sentence Encoder [1] with weights provided through TensorFlow Hub[7]. Disregarding visual feature aggregation, our visual and textual embedding networks both consist of the same network architecture: a fully connected layer with 1024 units and ReLU activation followed by a dropout layer with rate 0.2 and another fully connected layer with 256 units, which represent the final embedding space. To arrive at the final co-embedding, the output is normalized to the unit hypersphere.

Similar to [3], we perform the training using a bi-directional, pairwise triplet hard loss as seen in Equation 1, where v is a visual sample, s is a corresponding textual sample, v' and s' are non-matching visual and textual samples, d is the Euclidean distance in the co-embedding space, and α is a configurable margin parameter.

$$l_{PTH}(v,s) = \max_{s'} \left[\alpha + d(v,s) - d(v,s')\right]_+ + \max_{v'} \left[\alpha + d(v,s) - d(v',s)\right]_+ \quad (1)$$

During training, only the visual and textual embedding networks are trained; the pretrained visual and textual feature encoders remain fixed to greatly reduce required training resources. The networks are trained on a mixture of captioned

[6] https://keras.io/api/applications/inceptionresnetv2
[7] https://tfhub.dev/google/universal-sentence-encoder/4

video and image datasets consisting of Flickr30k [19], Microsoft COCO [9], MSR-VTT [18], TextCaps [13], TGIF [8], and VaTeX [17].

4 VR Query Formulation

Based on our experience from the previous instance of the VBS, we integrated a offline speech-to-text solution utilizing Mozilla DeepSpeech[8] and implemented interfaces to make the existing temporal query modalities and new pose query modalities accessible from VR.

Temporal Queries: Previously, vitrivr-VR did not allow the specification of temporal context during query formulation. Utilizing the same temporal query logic recently added to the vitrivr stack [5] , we have implemented a temporal query interface in VR. It allows the specification of temporal context of query terms by grabbing and reordering them in 3D space.

Pose Queries: Leveraging the intuitive 3D manipulation options in VR, vitrivr-VR supports a new query-by-pose feature. To formulate a pose query, a life size, mannequin-like pose representation can be manipulated by grabbing and rotating joints. The adjusted pose representation is framed using a virtual camera and projected onto a canvas to create a pose query.

5 Conclusion

In this paper we present the state of the vitrivr-VR system, with which we plan to participate in the VBS 2022. We describe our system, our visual-text co-embedding feature, changes to our speech-to-text method, and our new temporal and pose query formulation interfaces.

Acknowledgements

This work was partly supported by the Swiss National Science Foundation (project "Participatory Knowledge Practices in Analog and Digital Image Archives", contract no. CRSII5_193788).

References

1. Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., Kurzweil, R.: Universal sentence encoder. CoRR (2018)
2. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition (2009)

[8] https://github.com/mozilla/DeepSpeech

3. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives. In: British Machine Vision Conference 2018 (2018)
4. 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 (2020)
5. 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: International Conference on Multimedia Modeling (2022)
6. Heller, S., Sauter, L., Schuldt, H., Rossetto, L.: Multi-stage queries and temporal scoring in vitrivr. In: International Conference on Multimedia & Expo Workshops (2020)
7. Li, X., Xu, C., Yang, G., Chen, Z., Dong, J.: W2VV++: fully deep learning for ad-hoc video search. In: International Conference on Multimedia (2019)
8. Li, Y., Song, Y., Cao, L., Tetreault, J.R., Goldberg, L., Jaimes, A., Luo, J.: TGIF: A new dataset and benchmark on animated GIF description. In: Conference on Computer Vision and Pattern Recognition (2016)
9. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: European Conference on Computer Vision (2014)
10. 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., Jónsson, B.t.: Is the reign of interactive search eternal? findings from the video browser showdown 2020. ACM TOMM (2021)
11. Rossetto, L., Giangreco, I., Schuldt, H.: Cineast: A multi-feature sketch-based video retrieval engine. In: IEEE International Symposium on Multimedia (2014)
12. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C - A research video collection. In: International Conference on Multimedia Modeling (2019)
13. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: A dataset for image captioning with reading comprehension. In: European Conference on Computer Vision (2020)
14. 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 (2021)
15. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI Conference on Artificial Intelligence (2017)
16. Tran, L., Nguyen, M., Nguyen, T., Healy, G., Caputo, A., Nguyen, B.T., Gurrin, C.: A VR interface for browsing visual spaces at VBS2021. In: International Conference on Multimedia Modeling (2021)
17. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: International Conference on Computer Vision (2019)
18. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: Conference on Computer Vision and Pattern Recognition (2016)
19. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguistics (2014)