Multi-Modal Interactive Video Retrieval with Temporal Queries

Silvan Heller[1][0000−0001−5386−330X], Rahel Arnold[1][0000−0002−5881−4432],
Ralph Gasser[1][0000−0002−3016−1396], Viktor Gsteiger[1][0000−0002−6750−5500],
Mahnaz Parian-Scherb[1][0000−0001−7063−8585],
Luca Rossetto[2][0000−0002−5389−9465], Loris Sauter[1][0000−0001−8046−0362],
Florian Spiess[1][0000−0002−3396−1516], and Heiko Schuldt[1][0000−0001−9865−6371]

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

Abstract. This paper presents the version of vitrivr participating at the Video Browser Showdown (VBS) 2022. vitrivr already supports a wide range of query modalities, such as color and semantic sketches, OCR, ASR and text embedding. In this paper, we briefly introduce the system, then describe our new approach to queries specifying temporal context, ideas for color-based sketches in a competitive retrieval setting and a novel approach to pose-based queries.

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

1 Introduction

The Video Browser Showdown (VBS) is an annual evaluation campaign for interactive video retrieval systems [9], with its 11^{th} iteration happening in 2022. At VBS, system operators are tasked with finding relevant items within a large video collection. In the 2022 installment – for the first time – both the first and second shard of the V3C Dataset [13] are being used, resulting in approximately 2'300 hours of video. vitrivr is an open-source, multi-modal multimedia retrieval system which has participated in the VBS several times, including last year's installment [5]. A plethora of query modalities are supported by the system [4], notable in a competitive setting are color and semantic sketches, OCR, ASR, and text embedding [14]. The retrieval model is discussed in [7], and details on the storage layer can be found in [3]. Built as a general-purpose tool for multimedia retrieval, vitrivr is also used in other contexts such as lifelog retrieval [6].

In this paper, we present the version of vitrivr we plan to participate with at VBS 2022. Our focus for this participation lies on the improvements we have made to temporal queries, new ideas for making color sketches useful in a competitive setting and a novel approach to pose-based queries.

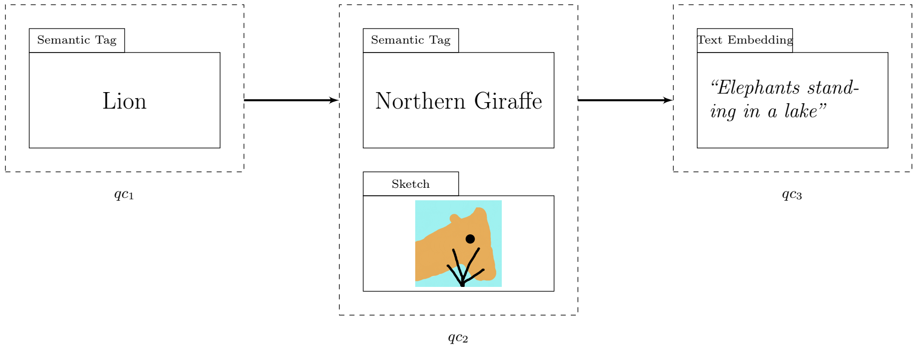

Fig. 1. A multi-modal temporal query consisting of three query containers (dashed, qc_1, qc_2 and qc_3), each with different modalities. This is a formulated query for "Shots of first a lion (semantic tag), followed by a giraffe (semantic tag) whose head is visible while eating a branch (sketch) concluded by a shot of *Elephants standing in a lake* (text embedding)". The temporal order is given through the order of query containers.

The remainder of this paper is structured as follows: Section 2 shows our algorithm for temporal queries, Section 3 introduces new ideas for color sketches, Section 4 discusses our approach to pose-based queries, and Section 5 concludes.

2 Temporal Queries

A temporal query as illustrated in Figure 1 consists of multiple, ordered similarity sub-queries of various and possibly different modalities and time distances between the sub-queries. The results of the sub-queries are then aggregated and scored according to the order of the query containers and the specified time distances. For notation purposes, we define a video V which is segmented into a list of segments $S = \langle s_1, s_2, ...s_m \rangle$. Given a temporal query TQ and its list of query containers $TQ = \langle qc_1, ..., qc_n \rangle$, with each query container being a possibly multi-modal query, each query container is executed seperately. The result is then a set RS of scored segments per container, i.e. $RS_i = \langle \langle s_f, s_g, ..., s_l \rangle, qc_i \rangle$, with a segment possibly appearing in multiple elements of the result set if it is a suitable result for multiple containers.

The aggregation process creates sequences of segments where sequences with more and higher scored result segments from multiple query containers are preferred for each object. A temporal sequence TS is defined as an ordered list of segment-container tuples $TS = \langle \langle s_i, qc_a \rangle, \langle s_j, qc_b \rangle, \ldots, \langle s_z, qc_n \rangle \rangle$. We have experimented with different algorithms, and will describe the current scoring algorithm utilised by vitrivr here:

1. For each tuple $\langle s_i, qc_a \rangle$ within the flattened list of results, temporal sequence candidates are constructed. Segments are appended to a temporal sequence if they follow the ordering specified by the user, i.e., for $TS = \langle s_1, qc_i \rangle$, only tuples $\langle s_x, qc_j \rangle$ with $j > i$ are considered.

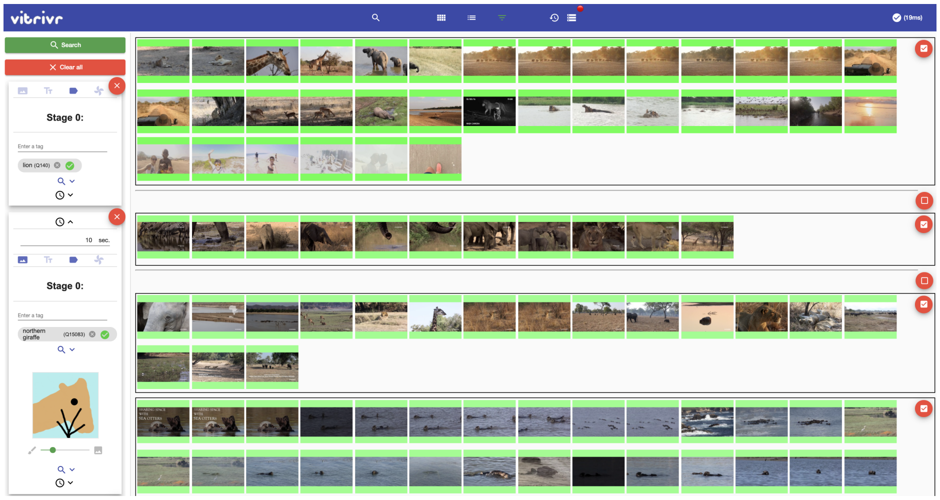

Fig. 2. The improved Temporal Scoring View of vitrivr with a temporal distance query and two toggled results.

2. Scores are aggregated within a sequence TS (with the first element $\langle s_i, qc_a\rangle$) with a decay function. For a given time difference t between two elements, all segments except the first ($TS \setminus \langle s_i, qc_a\rangle$) have their score multiplied with $adj(t) = e^{-|l\cdot(t-m)|}$, with l being the penalty of being not at the user-specified distance (we use $l = 0.1$), and $m \leq 0$ being the time to the next segment as defined by the temporal query. Items that are within the perfect time distance receive no penalty on their score.
3. After normalizing the score with respect to the number of sub-queries (3 in our example), the highest-scoring sequences are selected.

In Figure 2, we show a screenshot of vitrivr's user interface. The query containers are on the left, with the additional temporal distance between the first and the second query container. Additionally, results are grouped by video and their visibility can be interactively toggled by the user if they do not fulfill the search criteria. The evaluation [11] of the VBS competition in 2020 showed that frequent use of temporal queries increases the performance of retrieval systems. Different systems at VBS employ different scoring and aggregation algorithms. In the VIRET system [8], temporal aggregation is evaluated for each modality separately. VIREO [10] displays two canvases to the user to input two object-sketch queries at the timestamp t and t' with $t < t'$. The similarity is then determined between a video and the two queries. SomHunter [15] uses a fixed number of three adjacent temporal queries. Visione also introduced temporal querying for the 2021 version of their retrieval system [1]. They employ a similar strategy to the VIRET system with the possibility to provide two queries that should occur within a certain threshold of time.

3 Color Sketches

Up until now, vitrivr's retrieval engine Cineast [12] uses various color features, such as aggregation based features (histogram, median, and average color), and the Color Layout Descriptor among other features for visual similarity. These features are grouped into categories with empirically determined weights.

When differentiating between colors, care must also always be taken to ensure that the different nuances are visible to the human eye. This is particularly important during query formulation. We (fine-) tune color features used in vitrivr's retrieval engine based on a systematic analysis of the currently used approaches. We expect color sketches in combination with staged querying [7] to improve query formulation. Additionally, we aim on bridging the gap between perceived color input and actual extraction, potentially limiting the color palette drastically.

4 Pose-based Queries

As an addition to vitrivr, we include a pose-based query mode to search and find human-specific poses using the detailed location of detected body parts. The new feature module captures joint information, which is provided by the human 2D pose estimation framework OpenPose [2]. This algorithm localizes human body parts such as shoulders, ankles and knees from video frames. We use the 18 key-points model and the joint location coordinates together with the part affinity field information and store it in a 3 channel feature vector, which is used for finding specific poses.

On the user interface side, a new mode for query formulation based on pose is added. This new mode allows the user to describe the pose they are searching for using a canvas with a skeleton consisting of 18 key-points. Individual joints can be moved to specify the approximate pose of a person. To simplify lookup for common poses, a selection of preset configurations is provided.

During retrieval, the coordinates of the query are compared to stored features in the database. To measure the similarity, we use normalized coordinates to cover the instances in different locations of the frame. Normalization is done by determining the center of mass for each identified skeleton and then moving the origin of the coordinate system to that center of mass. In addition to the coordinates, we also consider the distances between joints to minimize the perspective effects.

5 Conclusion

In this paper, we presented the version of vitrivr with which we plan to participate at VBS 2022. Both temporal queries and staged querying using color sketches have shown promising results in past iterations of VBS, and pose-based queries as yet another modality could be interesting for challenging Visual KIS tasks or AVS tasks.

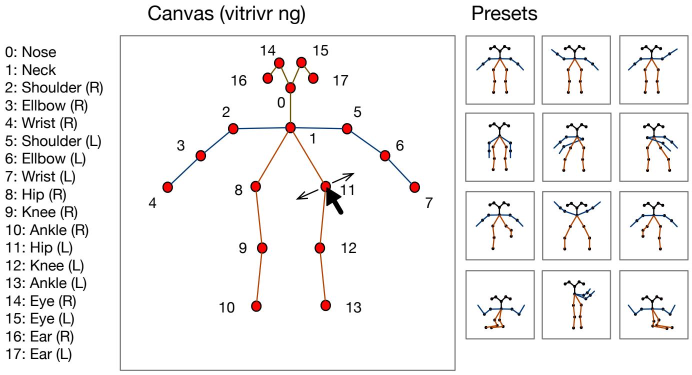

Fig. 3. Sketching mechanism for posed-based queries. Users can either adjust the skeleton's joints individually using their mouse or select a preset pose from a list.

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. Amato, G., Bolettieri, P., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: VISIONE at video browser showdown 2021. 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. 473–478. Springer (2021)
2. Cao, Z., Hidalgo, G., Simon, T., Wei, S., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2021)
3. Gasser, R., Rossetto, L., Heller, S., Schuldt, H.: Cottontail DB: an open source database system for multimedia retrieval and analysis. In: MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. pp. 4465–4468. ACM (2020)
4. Gasser, R., Rossetto, L., Schuldt, H.: Multimodal multimedia retrieval with vitrivr. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 2019. pp. 391–394. ACM (2019)
5. 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 - 27th International Conference, MMM 2021, Prague, Czech Republic, June 22-24, 2021, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12573, pp. 435–440. Springer (2021)
6. Heller, S., Gasser, R., Parian-Scherb, M., Popovic, S., Rossetto, L., Sauter, L., Spiess, F., Schuldt, H.: Interactive multimodal lifelog retrieval with vitrivr at LSC

2021. In: Proceedings of the 4th Annual on Lifelog Search Challenge, LSC@ICMR 2021, Taipei, Taiwan, 21 August 2021. pp. 35–39. ACM (2021)

7. Heller, S., Sauter, L., Schuldt, H., Rossetto, L.: Multi-stage queries and temporal scoring in vitrivr. In: 2020 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2020, London, UK, July 6-10, 2020. pp. 1–5. IEEE (2020)
8. Lokoc, J., Kovalcík, G., Soucek, T., Moravec, J., Cech, P.: A framework for effective known-item search in video. In: Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019. pp. 1777–1785. ACM (2019)
9. 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 Trans. Multimedia Comput. Commun. Appl. **17**(3) (Jul 2021)
10. Nguyen, P.A., Wu, J., Ngo, C., Francis, D., Huet, B.: VIREO @ video browser showdown 2020. In: MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 11962, pp. 772–777. Springer (2020)
11. Rossetto, L., Gasser, R., Lokoc, J., Bailer, W., Schoeffmann, K., Münzer, B., Soucek, 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 Trans. Multim. **23**, 243–256 (2021)
12. Rossetto, L., Giangreco, I., Schuldt, H.: Cineast: A multi-feature sketch-based video retrieval engine. In: 2014 IEEE International Symposium on Multimedia, ISM 2014, Taichung, Taiwan, December 10-12, 2014. pp. 18–23. IEEE Computer Society (2014)
13. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C - A research video collection. In: MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11295, pp. 349–360. Springer (2019)
14. 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 - 28th International Conference, MMM 2022 Proceedings. Lecture Notes in Computer Science, Springer (2022)
15. Veselý, P., Mejzlík, F., Lokoc, J.: Somhunter V2 at video browser showdown 2021. 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. 461–466. Springer (2021)