Novice-Friendly Text-based Video Search with vitrivr

Loris Sauter
loris.sauter@unibas.ch
University of Basel
Basel, Switzerland

Heiko Schuldt
heiko.schuldt@unibas.ch
University of Basel
Basel, Switzerland

Raphael Waltenspül
raphael.waltenspuel@unibas.ch
University of Basel
Basel, Switzerland

Luca Rossetto
rossetto@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

ABSTRACT

Video retrieval still offers many challenges which can so far only be effectively mediated through interactive, human-in-the-loop retrieval approaches. The vitrivr multimedia retrieval stack offers a broad range of query mechanisms to enable users to perform such interactive retrieval. While these multiple mechanisms offer various options to experienced users, they can be difficult to use for novices. In this paper, we present a minimal user interface geared towards novice users that only exposes a subset of vitrivr's functionality but simplifies user interaction.

CCS CONCEPTS

• **Information systems** → **Search interfaces**; **Video search**.

KEYWORDS

Interactive Video Retrieval, Competitive Retrieval, Intuitive Retrieval

ACM Reference Format:
Loris Sauter, Heiko Schuldt, Raphael Waltenspül, and Luca Rossetto. 2023. Novice-Friendly Text-based Video Search with vitrivr. In *20th International Conference on Content-based Multimedia Indexing (CBMI 2023), September 20–22, 2023, Orléans, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3617233.3617262

1 INTRODUCTION

General-purpose video retrieval remains a difficult task that, in many instances, still requires an interactive, human-in-the-loop approach to work effectively. While tremendous progress has been made in the area of automatic content understanding and semantic representation of video content, driven by advances in deep-neural approaches, only comparatively little attention is paid to the interactive aspect of retrieval. The probably most influential catalyst for the development of interactive means for the retrieval of general-purpose video is the Video Browser Showdown (VBS) [7], which has provided a venue for the systematic evaluation of interactive video retrieval approaches for over a decade [19]. VBS offers a controlled setting in which participating teams solve a series of retrieval tasks at the same time and – in so far as external circumstances allow it – at the same place. The tasks to be solved are not previously known to the participants and target one or multiple segments of a pre-defined video dataset [17, 23]. Participants are scored by a central evaluation system [13] based on both accuracy and speed, which incentivizes efficient and effective retrieval approaches.

The vitrivr multimedia retrieval stack has been a long-time participant [15] to VBS. Originally designed exclusively as a video retrieval system [16] with a focus on largely visual and sketch-based query mechanisms, it has grown over the years to encompass many more media types [3, 4] and query modes [5, 8, 12, 14, 18, 20, 21]. Since many of these various capabilities are available via the same user interface, it can be difficult for inexperienced users to use it to the best of its abilities and to exploit all its features.

In this paper, we present a different user interface to the vitrivr stack which we call $vitrivr_{min}$. In contrast to the more generally utilized user interface *vitrivr-ng*, the new $vitrivr_{min}$ interface is optimized for VBS-like competitive settings and offers only the subset of functionality that has been found to be most useful in this particular setting. Specifically, it only uses textual input for query formulation and omits all other querying means apart from some limited relevance feedback.

The remainder of this paper is structured as follows: Section 2 focuses on the different ways vitrivr can query video content based on textual input. Section 3 discusses the various other query capabilities made available by vitrivr to its operators during the most recent instance of the VBS. Section 4 contrasts this to the minimalist approach exhibited by $vitrivr_{min}$ and discusses the advantages and drawbacks concerning a VBS-like setting. Section 5 concludes.

2 TEXT-BASED VIDEO SEARCH IN VITRIVR

vitrivr offers several means of querying for specific information contained in video via textual query representations.

Text on Screen. A comparatively lossless way of describing information in video via textual means is by targeting information that was of textual nature in the first place, such as actual text shown within a video. To extract such text, vitrivr uses an optical character recognition method optimized for video [22] and stores all the extracted text jointly for every video segment. Querying is then performed via fuzzy full-text search provided by vitrivr's underlying database engine Cottontail DB [2].

CBMI 2023, September 20–22, 2023, Orléans, France
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0912-8/23/09.
https://doi.org/10.1145/3617233.3617262

Speech. Spoken dialogue is present in a large fraction of the larger of the two benchmark datasets [11] and – assuming the issuer of the query speaks the language present in the video – offers an information source for which queries can be easily expressed. vitrivr offers means to query for such dialogue via full-text search in time-aligned video subtitles. For videos where no subtitles were available, we use the 'whisper' [10] speech transcription system to generate them.

Scene description. vitrivr has multiple means for enabling the search of semantic content within a video via its textual description. Initially built on full-text search in automatically generated scene captions, more recent methods employ multi-modal visual/text co-embeddings that produce vectors comparable via kNN search. Several such embedding methods are available in vitrivr, both custom-developed ones [6] as well as multiple implementations [1] of CLIP [9] using freely available pre-trained models.

3 VITRIVR AT VBS'23

For VBS'23, vitrivr was primarily focused on text-based queries, such as the ones introduced in Section 2. The retrieval stack does, however, also support a plethora of other query modalities. Specifically, these include –but are not limited to– the following:

Query-by-Example (QbE) operates on the proverb "*an image says more than a thousand words*", vitrivr empowers users to provide an example and search for visually similar elements in the collection.

Query-by-Sketch (QbS) works similarly to QbE. However, users are responsible to sketch their information need by hand. Additionally, they could also use an example image and enhance it by superimposing sketches on it.

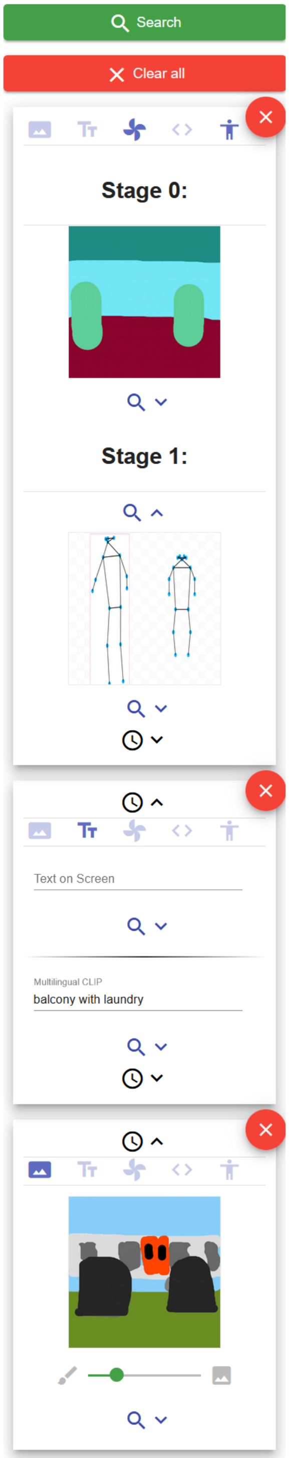
Query-by-SemanticSketch (QbSS) While visual similarity might be beneficial if available, the QbSS modality operates on semantic similarity. Users localize specific semantic concepts on a canvas and thus provide information on where a certain concept should be spatially located in a frame [14].

Query-by-Pose (QbP) Introduced in [6], it enables users to specify the information need as one or more posed human skeletons.

More-Like-This (MLT) At times, vitrivr might provide results that are close enough but not exactly what the users were looking for. MLT queries are used on an existing result which is then taken as input for a subsequent query.

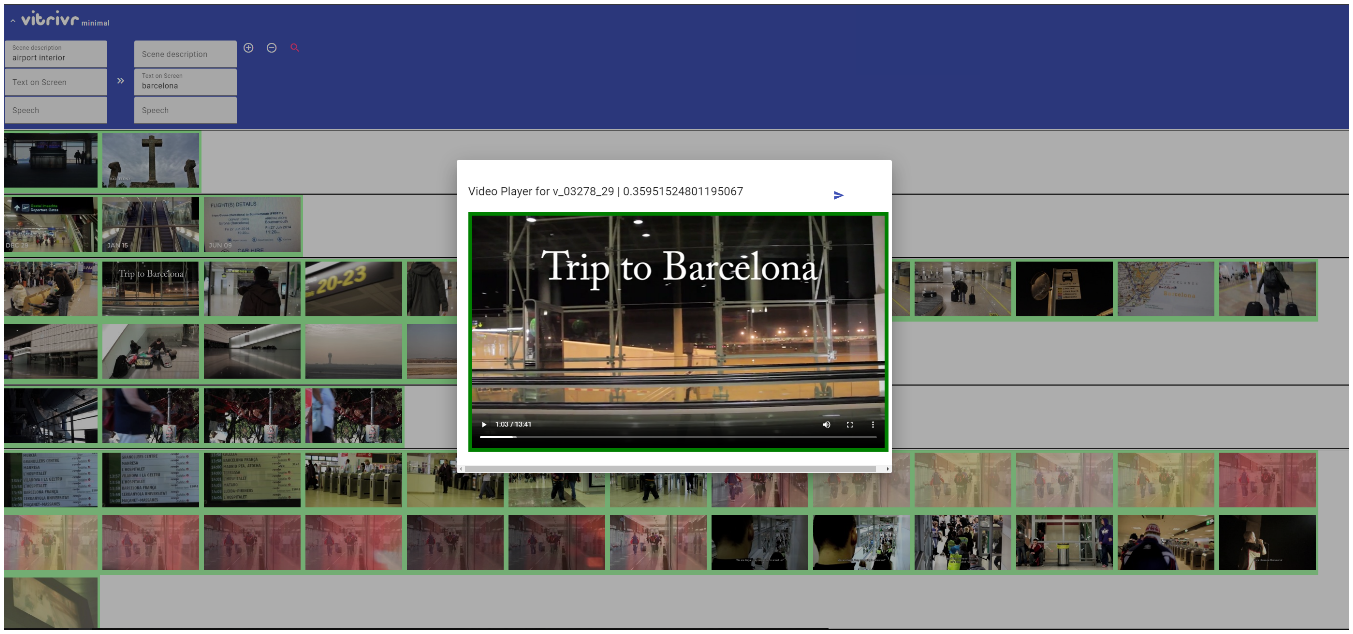
In addition to the query modalities, vitrivr provides various paths to express multiple requirements within the information need. First and foremost, multiple elements in different modalities can be specified, e.g., a query can be formulated by jointly providing a semantic sketch (QbSS modality) and a pose (QbP modality). vitrivr then presents results to the user that satisfy both query elements. To expand the expressiveness of queries even more, two additional contextualizations are possible.

On one hand, so-called *staged queries* represent a strong notion of the informal "*this* and also *that*", e.g., given in a first stage an upright pose and in a second stage an example image of meadows; this leads vitrivr to first execute a query for the upright pose and *on this result set*, the second query, the one with meadows, is executed.



Multilingual CLIP
balcony with laundry

Figure 1: Example of a query expressed in vitrivr-ng. The query targets three successive video sequences, the first described via a semantic sketch, further filtered by a pose query, the second described textually, and the third described via a visual sketch.



vitrivr minimal

Scene description: airport interior | Text on Screen | Speech

»

Scene description | Text on Screen: barcelona | Speech

Video Player for v_03278_29 | 0.35951524801195067

Figure 2: A screenshot of an example query in vitrivr$_{min}$. The query in the upper part represents "A scene that contains the semantic concept *airport interior* and subsequently there is the text *Barcelona* on screen". The video player was opened on the second thumbnail of the third row. This is indicated by the green border of the video player. The screenshot depicts five videos (divided by horizontal lines) with varying numbers of segments. The video player is open with video 3278 at 1:03 from [17]. A full demo video is available via https://youtu.be/F8fD0TaGt4Q

This results in a strong relationship between the second to the first stage.

On the other hand, an information need might as well contain a temporal aspect, such as "*first this, followed by that*", which can be expressed as a temporal query [8].

Figure 1 shows an example of how these different query formulation mechanisms could be used in conjunction for a textual known item search (KIS) task from VBS'23 that used the following description: *"A sequence of three shots: two people and a wall with posters, a balcony with laundry hanging on a rope and a train passing behind two standing tombstones. [...]"*.

4 NOVICE-FRIENDLY VIDEO SEARCH

In the 2021 installment of VBS, the top three systems all heavily relied on textual query modalities [7], some directly or indirectly influenced by the rise of joint visual text co-embedding approaches such as CLIP, hence text remains the dominant modality for video search. This trend towards the reliance on such visual/text co-embedding methods could also be observed in the 2022 and 2023 editions of VBS. We argue that text as an input modality is omnipresent and, thus, also the most novice-friendly approach to video search.

Query Formulation. Commercial products such as large video platforms (e.g., YouTube, Vimeo) provide little more than text input for users to search in video collections. Motivated by the above-mentioned analysis and commercial orientation, we removed every query modality except for the text-based ones and present operators with the means to express a query in three domains; (i) *Scene description*, (ii) *Text on Screen*, and (iii) *Speech*. We purposely do neither use the technical terms Optical Character Recognition (OCR) for "Text on Screen" nor Automated Speech Recognition (ASR) for "Speech", as the aim of this user interface is to be as intuitive as possible. In contrast to vitrivr-ng, where temporal queries are formulated vertically, multiple temporal terms are formulated horizontally, each with an icon in between indicating that time has passed. The entire query formulation region is located in the top bar, as opposed to a sidebar in vitrivr-ng.

Result Presentation. While vitrivr-ng provides various means of result presentation, in vitrivr$_{min}$, we limit the result presentation to a single view: The resulting segments are grouped by their source video and ranked by their merged score. Within each video, the keyframes of the segments are shown, and a visual indicator of the score is displayed. This indication of the score is similar to vitrivr-ng: Thumbnails have a green border representing the score s: $(\mathrm{r, g, b}) = (s', 255, s')$ with $s' = \lfloor (1 - s) \cdot 255 \rfloor$, i.e., the greener the higher the score. The videos are sorted by their score, as in a traditional result list, with the item with the highest score being the topmost one. Since only segments are scored, a video's score is the maximum score of its segments.

When hovering over a preview image, two buttons are shown. One enables the user to directly start a new *More-Like-This* query for segments similar to an already retrieved result. The other enables operators to quick-submit to the evaluation server [13]. Clicking on a preview image opens an overlay with a video player.

Browsing. While browsing through the results is enabled, in addition to this, vitrivr$_{min}$ empowers users to closely inspect ranked segments within their video context through a traditional video player with all controls enabled, as shown in Figure 2. Giving users a fully controllable video player enables quick in-video browsing through the progress bar or various playback speeds. To highlight which portion of the played video is the selected segment, it is visually indicated by a dark-green border around the video player. To further explore the video collection, users can move onward from any retrieved segment with a *More-Like-This* query to inspect semantically and visually similar results.

One of the advantages of this minimalist user interface is that, particularly for competition-like settings, there is almost no learning curve for its operation. Operators are presented with exactly three modalities to formulate their query with, all using natural language and labeled with commonly known terms (“Scene description”, “Text on Screen”, “Speech”).

In addition to that, upon hovering over a result tile, a *More-Like-This* query button, as well as a submit button, are available, both using often-used icons as well as tooltips. However, as simple as the user interface is, formulating prompts for approaches based on visual text co-embedding might not be trivial and still require some time to get familiar with. In particular, humans were trained over the last decades to use as little textual information as possible and to keep terms general in order to find as many results as possible on popular (web) search engines such as Google or DuckDuckGo. Prompts for our visual text co-embedding appear to yield better results when using proper sentences. This behavior might not be intuitive for novices. Furthermore, vitrivr$_{min}$ does not provide any information on the collection to be searched in. Thus, operators have to use their imagination or the competition-given information need in order to formulate queries, in contrast to systems that provide automatic text completion for query suggestions. In the context of the CBMI special session, we aim to gain insights into the effectiveness of the vitrivr$_{min}$ user interface in comparison to the existing, more complex one.

5 CONCLUSION & OUTLOOK

In this paper, we present *vitrivr$_{min}$*, a minimalist user interface for the open-source content-based multimedia retrieval system vitrivr. Stripped off of anything that might obstruct novice users, we rely on textual input only, which recent studies have shown to be very effective. In particular, vitrivr$_{min}$ empowers users to formulate queries that represent (i) *Speech*, (ii) *Text on Screen*, and (iii) *Scene descriptions*. Additionally, users are able to inspect resulting segments in their context, i.e., their source video. In order to support a competition-style evaluation based on the annual Video Browser Showdown, vitrivr$_{min}$ provides the means to intuitively submit retrieved results to the evaluation server. Complementing VBS, we expect novice-centric evaluations, such as the one held at CBMI’23, to yield new insights into the effectiveness of our minimalist approach. A full demo video of vitrivr$_{min}$in action is available via https://youtu.be/F8fD0TaGt4Q.

ACKNOWLEDGMENTS

This work was partly supported by the Swiss National Science Foundation through project “MediaGraph” (contract no. 202125) and by InnoSuisse through project “Archipanion” (contract no. 104.460 IP-ICT).

REFERENCES

[1] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. Reproducible scaling laws for contrastive language-image learning. *CoRR* abs/2212.07143 (2022). https://doi.org/10.48550/arXiv.2212.07143 arXiv:2212.07143

[2] Ralph Gasser, Luca Rossetto, Silvan Heller, and Heiko Schuldt. 2020. 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.* ACM, 4465–4468. https://doi.org/10.1145/3394171.3414538

[3] Ralph Gasser, Luca Rossetto, and Heiko Schuldt. 2019. 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.* ACM, 391–394. https://doi.org/10.1145/3323873.3326921

[4] Ralph Gasser, Luca Rossetto, and Heiko Schuldt. 2019. Towards an All-Purpose Content-Based Multimedia Information Retrieval System. *CoRR* abs/1902.03878 (2019). arXiv:1902.03878 http://arxiv.org/abs/1902.03878

[5] Prateek Goel, Ivan Giangreco, Luca Rossetto, Claudiu Tanase, and Heiko Schuldt. 2017. "Hey, vitrivr!" - A Multimodal UI for Video Retrieval. In *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings (Lecture Notes in Computer Science, Vol. 10193).* 749–752. https://doi.org/10.1007/978-3-319-56608-5_75

[6] Silvan Heller, Rahel Arnold, Ralph Gasser, Viktor Gsteiger, Mahnaz Parian-Scherb, Luca Rossetto, Loris Sauter, Florian Spiess, and Heiko Schuldt. 2022. Multi-modal Interactive Video Retrieval with Temporal Queries. In *MultiMedia Modeling - 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6-10, 2022, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13142).* Springer, 493–498. https://doi.org/10.1007/978-3-030-98355-0_44

[7] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoc, Andreas Leibetseder, Frantisek Mejzlík, Ladislav Peska, Luca Rossetto, Konstantin Schall, Klaus Schoeffmann, Heiko Schuldt, Florian Spiess, Ly-Duyen Tran, Lucia Vadicamo, Patrik Veselý, Stefanos Vrochidis, and Jiaxin Wu. 2022. Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown. *Int. J. Multim. Inf. Retr.* 11, 1 (2022), 1–18. https://doi.org/10.1007/s13735-021-00225-2

[8] Silvan Heller, Loris Sauter, Heiko Schuldt, and Luca Rossetto. 2020. 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.* IEEE, 1–5. https://doi.org/10.1109/ICMEW46912.2020.9105954

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[10] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *CoRR* abs/2212.04356 (2022). https://doi.org/10.48550/arXiv.2212.04356 arXiv:2212.04356

[11] Luca Rossetto. 2022. You were saying? - Spoken Language in the V3C Dataset. *CoRR* abs/2212.07835 (2022). https://doi.org/10.48550/arXiv.2212.07835 arXiv:2212.07835

[12] Luca Rossetto, Ralph Gasser, Silvan Heller, Mahnaz Amiri Parian, and Heiko Schuldt. 2019. Retrieval of Structured and Unstructured Data with vitrivr. In *Proceedings of the ACM Workshop on Lifelog Search Challenge, LSC@ICMR 2019, Ottawa, ON, Canada, 10 June 2019.* ACM, 27–31. https://doi.org/10.1145/3326460.3329160

[13] Luca Rossetto, Ralph Gasser, Loris Sauter, Abraham Bernstein, and Heiko Schuldt. 2021. 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).* Springer, 385–390. https://doi.org/10.1007/978-3-030-67835-7_33

[14] Luca Rossetto, Ralph Gasser, and Heiko Schuldt. 2019. Query by Semantic Sketch. *CoRR* abs/1909.12526 (2019). arXiv:1909.12526 http://arxiv.org/abs/1909.12526

[15] Luca Rossetto, Ivan Giangreco, Heiko Schuldt, Stéphane Dupont, Omar Seddati, T. Metin Sezgin, and Yusuf Sahillioglu. 2015. IMOTION - A Content-Based Video

Retrieval Engine. In *MultiMedia Modeling - 21st International Conference, MMM 2015, Sydney, NSW, Australia, January 5-7, 2015, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 8936)*. Springer, 255–260. https://doi.org/10.1007/978-3-319-14442-9_24

[16] Luca Rossetto, Ivan Giangreco, Claudiu Tanase, and Heiko Schuldt. 2016. vitrivr: A Flexible Retrieval Stack Supporting Multiple Query Modes for Searching in Multimedia Collections. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*. ACM, 1183–1186. https://doi.org/10.1145/2964284.2973797

[17] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A. Butt. 2019. 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)*. Springer, 349–360. https://doi.org/10.1007/978-3-030-05710-7_29

[18] Loris Sauter, Luca Rossetto, and Heiko Schuldt. 2018. Exploring Cultural Heritage in Augmented Reality with GoFind!. In *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality, AIVR 2018, Taichung, Taiwan, December 10-12, 2018*. IEEE Computer Society, 187–188. https://doi.org/10.1109/AIVR.2018.00041

[19] Klaus Schoeffmann, Jakub Lokoc, and Werner Bailer. 2020. 10 years of video browser showdown. In *MMAsia 2020: ACM Multimedia Asia, Virtual Event / Singapore, 7-9 March, 2021*. ACM, 73:1–73:3. https://doi.org/10.1145/3444685.3450215

[20] Florian Spiess, Ralph Gasser, Silvan Heller, Luca Rossetto, Loris Sauter, and Heiko Schuldt. 2021. Competitive Interactive Video Retrieval in Virtual Reality with vitrivr-VR. 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)*. Springer, 441–447. https://doi.org/10.1007/978-3-030-67835-7_42

[21] Claudiu Tanase, Ivan Giangreco, Luca Rossetto, Heiko Schuldt, Omar Seddati, Stéphane Dupont, Ozan Can Altiok, and T. Metin Sezgin. 2016. Semantic Sketch-Based Video Retrieval with Autocompletion. In *Companion Publication of the 21st International Conference on Intelligent User Interfaces, IUI 2016, Sonoma, CA, USA, March 7-10, 2016*. ACM, 97–101. https://doi.org/10.1145/2876456.2879473

[22] Alexander Theus, Luca Rossetto, and Abraham Bernstein. 2022. HyText - A Scene-Text Extraction Method for Video Retrieval. In *MultiMedia Modeling - 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6-10, 2022, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13142)*. Springer, 182–193. https://doi.org/10.1007/978-3-030-98355-0_16

[23] Quang-Trung Truong, Tuan-Anh Vu, Tan-Sang Ha, Jakub Lokoc, Yue Him Wong Tim, Ajay Joneja, and Sai-Kit Yeung. 2023. 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, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13833)*. Springer, 539–550. https://doi.org/10.1007/978-3-031-27077-2_42