A Comparison of Video Browsing Performance between Desktop and Virtual Reality Interfaces

Florian Spiess
University of Basel
Basel, Switzerland
florian.spiess@unibas.ch

Ralph Gasser
University of Basel
Basel, Switzerland
ralph.gasser@unibas.ch

Silvan Heller
University of Basel
Basel, Switzerland
silvan.heller@unibas.ch

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

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

ABSTRACT

Interactive retrieval with user-friendly and performant interfaces remains a necessity for video retrieval, even in light of significant gains in retrieval performance through multi-modal encoders. In recent years, novel interaction modalities such as virtual reality (VR) and augmented reality (AR) have gained popularity, but the best way to adapt paradigms from traditional retrieval interfaces, especially for result browsing and interaction, remains an open research question. In this paper, we compare two video retrieval interfaces in a controlled setting to gain insight into the differences in video browsing between VR and desktop interfaces. We formulate hypotheses explaining why there might be performance differences between the two interfaces, define metrics to test the hypotheses, and show results based on data gathered at an evaluation campaign. Our results show that VR interfaces can be competitive in browsing performance and indicate that there can even be an advantage when browsing larger result sets in VR.

CCS CONCEPTS

• **Information systems → Image search**; **Search interfaces**; **Query representation**; **Collaborative search**; • **Human-centered computing → Virtual reality**.

KEYWORDS

Video Browsing, Retrieval Performance, Virtual Reality Interfaces

ACM Reference Format:
Florian Spiess, Ralph Gasser, Silvan Heller, Heiko Schuldt, and Luca Rossetto. 2023. A Comparison of Video Browsing Performance between Desktop and Virtual Reality Interfaces. In *International Conference on Multimedia Retrieval (ICMR '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3591106.3592292

1 INTRODUCTION

Content-based video retrieval has benefited tremendously in recent years from advances in the area of artificial neural networks. Especially multi-modal encoders, such as OpenAI's CLIP [6], have resulted in substantial improvements in retrieval accuracy when using textual queries to search for visual content. Despite these advances, sufficiently complex queries or scenes that are difficult to describe concisely in a short phrase still require interactive approaches with a human in the loop for effective retrieval and result browsing, especially for very large data collections. For these processes to be efficient, it is necessary for the user to be able to inspect a sufficiently large result set quickly and adjust the querying strategy accordingly.

The Video Browser Showdown (VBS) [3] is an annual benchmarking initiative with the aim of measuring the relative performance of interactive video retrieval systems in a controlled environment. It uses pre-defined datasets [9, 15] that are made available to participants beforehand as well as several types of retrieval tasks. The tasks, which need to be solved by all participants concurrently and in the least possible time, are not previously known. This allows for a comparison of the end-to-end retrieval performance, not only focusing on isolated features but also considering the interplay between user and system.

Most interactive video retrieval systems today still rely on desktop user interfaces, which present results on a screen. In recent years, however, there has been interest in alternative modes, e.g., using head-mounted virtual reality (VR) displays.

As a case study, we observe the performance of two interfaces: vitrivr [10] and vitrivr-VR [13]. Both interfaces share the same underlying database system [2] and query processing engine [8], but differ in the way a user interacts with them. While vitrivr uses a more traditional browser-based user interface, vitrivr-VR leverages a virtual reality space to let the user express search queries and interact with retrieved results. The presented comparison is based on logs sent to the evaluation server [7] used during the VBS complemented by locally collected information, and goes beyond the analysis of previous evaluations [4].

During the 2023 VBS, vitrivr-VR outperformed vitrivr. In this paper, we investigate the source of this difference in performance. We formulate three hypotheses (H1—H3) as to why there might be performance differences, and define four metrics (M1—M4) which can be used to test these hypotheses. These can be used to compare

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0178-8/23/06.
https://doi.org/10.1145/3591106.3592292

the performance of retrieval system interfaces and also serve as a starting point for future comparisons of desktop and virtual reality environments.

The remainder of this paper is structured as follows: Section 2 formulates the hypotheses and metrics, Section 3 outlines the specific interfaces, setting, and data we use for our analysis. Section 4 then presents the results of the comparison, and Sections 5 and 6 discuss their implications and limitations. Finally, Section 7 concludes.

2 COMPARING VR AND DESKTOP INTERFACE BROWSING

2.1 Hypotheses

To understand the observed difference in performance of two interfaces, we propose three hypotheses describing how these differences could have come about. To formulate these hypotheses, we identify three aspects of the retrieval process, where the systems might perform differently: the *query formulation time*, the *query quality*, and the *browsing time*. All three aspects potentially have a direct impact on the observed performance in the evaluation. Our hypotheses are as follows:

- **H1 – Query formulation time:** Queries are formulated more quickly by the operators of one type of interface, which can lead to more browsing time or faster query iteration.
- **H2 – Result quality:** Queries issued by the operators of one type of interface lead to results containing retrieval targets at lower ranks, making results easier to find.
- **H3 – Browsing performance:** Browsing performance of the operators of one type of interface is better. This can lead to target objects at similar ranks in the result sets being found more quickly, or reducing the number of times targets present in the result set were missed.

By analyzing the data in an attempt to substantiate or dismiss these three hypotheses, we can only expect to determine from which part of the interactive retrieval process the difference originates, and not directly if this difference originates from the inherent user interface design or the skill of the operators. Despite this caveat, we expect this analysis to provide useful insight to improve development of interactive retrieval systems.

2.2 Metrics

To test our hypotheses, we define the following metrics and indicate the hypothesis it will support:

- **M1 – Time to first results (H1):** We measure the time from the start of a task to the first result set appearance. Given that network conditions and the backend are identical for the VR and desktop interfaces, we can use this as a proxy to test differences in query formulation time.
- **M2 – Best rank (H2):** We determine the lowest rank of a target item achieved in the results of each query to measure the result quality. This is also a proxy for query quality under the assumption that a better query will cause target items to appear at lower ranks.
- **M3 – Browsing miss@k (H3):** We measure how many browsing misses, i.e., the correct item was in the result set but not submitted, occur per user and interface at or below a given rank k. This serves as a proxy for the browsing performance of the operator-display method combinations.
- **M4 – Relation between best rank & browsing time (H3):** We analyse the relation between the best rank, i.e., the first correct occurrence in the result set, and the time until the correct item was submitted. This indicates how fast the operators were able to find the correct result within a result set with a given display method.

3 EVALUATION PROCEDURE

3.1 Benchmark

The benchmark has two different task types: known-item search (KIS) tasks, where there is exactly one correct answer in the dataset, and ad-hoc video search (AVS), where there can be unlimited correct answers that match the query. For the former, there are three distinct sub-types: V-KIS uses V3C [9], a large dataset of diverse video content, to select a unique short video sequence of a few seconds in length. V-KIS M uses the same query type but on a different dataset [15] consisting of under-water videos with large amount of visual redundancy. T-KIS tasks again use the V3C dataset but rather than showing the actual video sequence in question, only a textual description is provided. Finally, AVS tasks again use V3C but use a brief textual description that might match an arbitrary number of video sequences.

3.2 Interfaces

vitrivr-VR is a virtual reality interface for immersive query formulation and results exploration. During the VBS, vitrivr-VR only used text-based features including a visual-text co-embedding [11], a multi-lingual OpenCLIP [1], on-screen text search, and automatic speech recognition. To facilitate text entry, vitrivr-VR provides speech-to-text and a virtual word-gesture keyboard [14]. Results exploration in vitrivr-VR makes use of the immersive virtual space in three main ways: a cylindrical grid-based results display that surrounds the user, a grabbable video player that can be placed anywhere in virtual space, and a multimedia drawer showing and allowing navigation through selected frames of a video.

vitrivr uses a more traditional browser-based approach for both query formulation and result presentation. While it offers a broad range of different querying methods, only the text-based queries were used in this experiment, analogously to what was used by vitrivr-VR. Results can be displayed either as a list of segments ranked by their similarity score or as a list of videos ordered by the highest score of any contained segment. Either organization enables the operator to play back the video at the relevant point in time by clicking the segment preview as well as to make submissions from any point in the selected video. During the VBS, primarily the second mode of result presentation was used.

3.3 Collected Data

The data used in this analysis was collected during the 2023 edition of the VBS, where 13 different retrieval systems participated, all solving the same tasks at the same time. The VBS uses an open-source evaluation server [7] that handles task presentation, collects submissions from participating teams, and facilitates the assessment of their correctness. It also offers the participants the option to

Table 1: Overall statistics per task type. Values for the Desktop modality are in the top half of each cell, values for the VR modality in the bottom.

Task Type	Points	Tasks solved	Correct Submissions	Incorrect Submissions
V-KIS	738	4 / 6	4	0
	877	**5** / 6	5	3
V-KIS M	618	4 / 6	4	1
	881	**6** / 6	6	0
T-KIS	**928**	**6** / 7	7[1]	1
	738	5 / 7	5	2
AVS	**702**	7 / 7	231	82
	699	7 / 7	236	109
Total	2986	21 / 26	246	84
	3195	**23** / 26	252	114

submit result logs, recording the video segments that were retrieved during their search activities. We use the data collected by the evaluation server, including these result logs and augment them with local logs to ensure completeness.

4 RESULTS

Table 1 shows an overview of the overall performance statistics of both modalities. The VR approach performs better overall in both score and number of tasks solved. Regarding different task types, vitrivr-VR performs better in both V-KIS categories, but vitrivr performs much better in T-KIS tasks, and marginally better in AVS tasks. The median time between task start and the first query results is shown in Table 2. For all task types, operators took longer to reach the first results in VR than on the desktop interface in the median. Since the backend and network conditions were identical between the two types of interfaces, we can assume that this reflects the query formulation time. The biggest difference was observed for V-KIS tasks, and the smallest for AVS tasks. In case of both interfaces the time to first results was longest for V-KIS tasks and shortest for T-KIS tasks.

Table 3 shows the median rank of both the targeted video segment as well as the first rank of any segment from the relevant video for all known-item search tasks. In cases where more than half of the result sets did not contain an exactly matching segment, the table shows *n/a* since the median is undefined in this case. It can be seen that the median rank of a relevant result was better when using the desktop modality, independent of task type.

Tasks for which no correct submission was made are shown in Table 4 grouped by task type and modality. While the target segment was not retrieved in all cases, for all listed tasks the target video was contained within the result set but missed by the operators. The table lists both the best rank for the video and the target segment, if it was contained in the result set. In the latter case, we consider this a browsing miss (M3).

Table 2: Median time from task start to first explorable results in seconds, grouped by task type and modality (M1).

Task Type	Modality	Median Time to First Result in Seconds	Difference
V-KIS	Desktop	**30.89**	+26.9%
	VR	39.20	
V-KIS M	Desktop	**27.43**	+17.2%
	VR	32.15	
T-KIS	Desktop	**21.29**	+24.7%
	VR	26.55	
AVS	Desktop	**25.41**	+11.98%
	VR	28.45	
Any	Desktop	**25.81**	+24.18%
	VR	32.05	

Table 3: Median rank of the first item in the result set that either matched the target or stemmed from the same video as the target, grouped by task type and modality (M2).

Task Type	Modality	Median Best Video Rank	Median Best Segment Rank
V-KIS	Desktop	**30.5**	n/a
	VR	34.5	n/a
V-KIS M	Desktop	**20**	20
	VR	71.5	74.5
T-KIS	Desktop	**2**	4
	VR	3	n/a
Any	Desktop	**7**	65
	VR	24	1201

Table 4: Best ranks of target video and segment for all tasks where no correct submission was made (M3), grouped by task type and modality.

Task Type	Modality	Misses	Min Video Ranks	Min Segment Ranks
V-KIS	Desktop	2	12, 202	n/a, n/a
	VR	**1**	24	n/a
V-KIS M	Desktop	2	24, 641	24, 1613
	VR	**0**	–	-
T-KIS	Desktop	**1**	2	n/a
	VR	2	14, 64	14, n/a

Figure 1 illustrates, for every correct submission, the rank at which the first result from the correct video appeared in the result set on the horizontal axis and the time taken until the operator identified and submitted the correct segment on the vertical axis. The lines indicate the best linear fit per modality. Two outlier datapoints with a rank over 200 and a browsing time over 2 minutes have been removed. Due to the small number of data points, the trend lines are shown for illustration purposes and should be considered as a rough indication rather than a concrete relation.

[1] One submission was manually set to be correct by a judge after a second correct submission was already received.

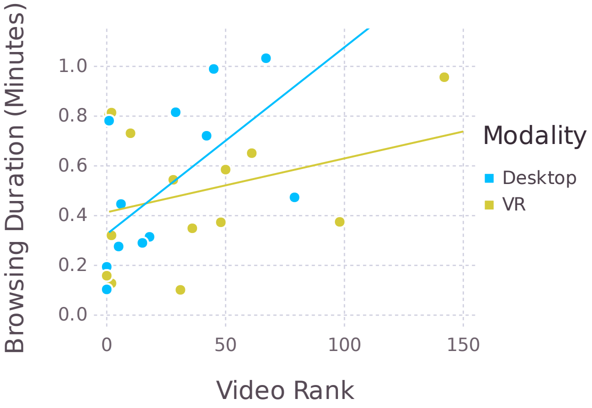

Figure 1: Browsing time vs. video rank for both interaction modalities (M4). Lines indicate best linear fit per modality.

5 DISCUSSION

The results presented in Table 2 allow us to rule out hypothesis H1, based on our assumption that time to first results is indicative of query formulation time. On the contrary, M1 shows that first results were available sooner for vitrivr than for vitrivr-VR. While this does not explain why vitrivr-VR performed better overall than vitrivr, it is an expected result of challenges to textual query formulation in VR [5].

For both modalities, first results were available the latest in the median for the two types of V-KIS tasks. V-KIS task cues consist of several seconds of video, which are only available once the task starts. Due to this, it is likely that operators only began formulating a query after having viewed the cue in its entirety. Since we are only able to measure the time from task start until first results, this initial viewing time is included for V-KIS tasks.

Table 3 shows that, in the median, targets were at lower ranks for query results of vitrivr than for vitrivr-VR. This data from M2 allows us to rule out hypothesis H2, as queries issued by operators of vitrivr-VR evidently did not lead to better results than those returned to operators of vitrivr. While in many cases the difference in median best rank between the interfaces is small, as would be expected of interfaces with access to the same retrieval features, in cases where they differ substantially vitrivr achieved the better results.

So far, we have been able to rule out hypotheses H1 and H2, and shown that operators in VR were slower to formulate their first query and overall did not issue better queries. As a consequence, the performance difference between the two interfaces must have occurred during the browsing phase of interactive retrieval. This is substantiated by the general trends shown in Table 4 and Figure 1. The results show that operators of vitrivr-VR were able to solve more tasks and where browsing misses did occur, they did not occur at significantly lower ranks than for vitrivr. Furthermore, although there is no clear trend, when the best ranks of targets were similar between the two interfaces, the tendency appears to be that operators of vitrivr-VR completed the task in less browsing time than the operators of vitrivr. As a result of these observations, we accept hypothesis H3 as the reason for the final performance difference.

In summary, the results of our analysis indicate that VR interfaces may perform worse for text-based query formulation, but could provide benefits for multimedia browsing.

6 LIMITATIONS

Our analysis is limited by a number of factors resulting from the design of the evaluation campaign and the data recorded by the interfaces. The main limitation is the fixed combination of two operators and interface instance pairs per team, as well as the relatively low number of tasks. While this does not limit our ability to determine during which part of the interactive retrieval process the performance differences arise, any performance differences could also be a result of the specific operators or an artefact of the low sample size.

While we are able to show that, in the median, first results were available later for vitrivr-VR than for vitrivr, our inference regarding query formulation time is only based on the reasonable assumption that backend and network, which were identical, had little impact. To be able to measure exactly how long initial query formulation takes, the two interfaces would need to record not only the time query results were returned, but also the time when the query was issued. Furthermore, accurately measuring the formulation time of further queries is a complex task, as it is difficult to determine when such query formulation starts and ends, since operators are able to browse existing results in parallel.

In this analysis we did not look at different stages of the video browsing process and instead treated everything after the return of the initial result set as a single step. As a result, we do not analyse differences within the browsing process, such as performance differences between inter- and intra-video browsing, and browsing strategies and behaviors.

7 CONCLUSION

In this paper, we presented a comparison between a desktop and a virtual reality-based video browsing approach in the context of interactive video retrieval. The analysis of the data collected during the VBS indicate, that while desktop-based approaches currently appear to be more effective in terms of query formulation, a VR-based approach can offer more effective results browsing. However, due to the conditions of the benchmark, we are so far unable to differentiate between the influence of the intrinsic properties of the interface and the performance of an individual system operator. Therefore, we need to treat the results in terms of the specific operator-interface combination and not in terms of the interfaces themselves. Nevertheless, the results hint at interesting insights that deserve to be investigated further.

In future work, we aim at repeating such an experiment with a larger number of system operators to obtain more reliable results, and to experiment with a hybrid desktop-VR approach to harness the advantages of both modalities [12].

ACKNOWLEDGMENTS

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).

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] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoč, Andreas Leibetseder, František Mejzlík, Ladislav Peška, 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. *International Journal of Multimedia Information Retrieval* 11, 1 (2022), 1–18. https://doi.org/10.1007/s13735-021-00225-2

[4] Silvan Heller, Florian Spiess, and Heiko Schuldt. 2023. A Tale of Two Interfaces: Vitrivr at the Lifelog Search Challenge. *Multimedia Tools and Applications* (2023), 1–25. https://doi.org/10.1007/s11042-023-15082-w

[5] Pascal Knierim, Thomas Kosch, Johannes Groschopp, and Albrecht Schmidt. 2020. Opportunities and Challenges of Text Input in Portable Virtual Reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, CHI 2020, Honolulu, HI, USA, April 25-30, 2020*. ACM, 1–8. https://doi.org/10.1145/3334480.3382920

[6] 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 *International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[7] Luca Rossetto, Ralph Gasser, Loris Sauter, Abraham Bernstein, and Heiko Schuldt. 2021. A System for Interactive Multimedia Retrieval Evaluations. In *MultiMedia Modeling (MMM) (Lecture Notes in Computer Science, Vol. 12573)*. Springer, 385–390.

[8] Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. 2014. Cineast: A Multi-feature Sketch-Based Video Retrieval Engine. In *International Symposium on Multimedia*. IEEE Computer Society, 18–23. https://doi.org/10.1109/ISM.2014.38

[9] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A. Butt. 2019. V3C – A Research Video Collection. In *MultiMedia Modeling*. Springer International Publishing, Cham, 349–360. https://doi.org/10.1007/978-3-030-05710-7_29

[10] Loris Sauter, Ralph Gasser, Silvan Heller, Luca Rossetto, Colin Saladin, Florian Spiess, and Heiko Schuldt. 2023. Exploring Effective Interactive Text-Based Video Search in vitrivr. In *MultiMedia Modeling - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13833)*, Duc-Tien Dang-Nguyen, Cathal Gurrin, Martha A. Larson, Alan F. Smeaton, Stevan Rudinac, Minh-Son Dao, Christoph Trattner, and Phoebe Chen (Eds.). Springer, 646–651. https://doi.org/10.1007/978-3-031-27077-2_53

[11] Florian Spiess, Ralph Gasser, Silvan Heller, Mahnaz Parian-Scherb, Luca Rossetto, Loris Sauter, and Heiko Schuldt. 2022. Multi-Modal Video Retrieval in Virtual Reality with Vitrivr-VR. In *MultiMedia Modeling*. Springer, 499–504. https://doi.org/10.1007/978-3-030-98355-0_45

[12] Florian Spiess, Ralph Gasser, Heiko Schuldt, and Luca Rossetto. 2023. The Best of Both Worlds: Lifelog Retrieval with a Desktop-Virtual Reality Hybrid System. In *Proceedings of the 2023 International Conference on Multimedia Retrieval* (Thessaloniki, Greece) *(ICMR '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3592573.3593107

[13] Florian Spiess, Silvan Heller, Luca Rossetto, Loris Sauter, Philipp Weber, and Heiko Schuldt. 2023. Traceable Asynchronous Workflows in Video Retrieval with vitrivr-VR. In *MultiMedia Modeling - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13833)*, Duc-Tien Dang-Nguyen, Cathal Gurrin, Martha A. Larson, Alan F. Smeaton, Stevan Rudinac, Minh-Son Dao, Christoph Trattner, and Phoebe Chen (Eds.). Springer, 622–627. https://doi.org/10.1007/978-3-031-27077-2_49

[14] Florian Spiess, Philipp Weber, and Heiko Schuldt. 2022. Direct Interaction Word-Gesture Text Input in Virtual Reality. In *International Conference on Artificial Intelligence and Virtual Reality*. IEEE, 140–144. https://doi.org/10.1109/AIVR56993.2022.00028

[15] 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