Interactive Video Retrieval Evaluation at a Distance

Comparing Sixteen Interactive Video Search Systems in a Remote Setting at the 10^{th} Video Browser Showdown

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 · Jiaxin Wu

Received: date / Accepted: date

S. Heller · V. Gsteiger · H. Schuldt · F. Spiess
Department of Mathematics and Computer Science
University of Basel, Basel, Switzerland
E-mail: silvan.heller@unibas.ch, v.gsteiger@gmail.com, heiko.schuldt@unibas.ch, florian.spiess@unibas.ch

W. Bailer
JOANNEUM RESEARCH, Graz, Austria
E-mail: werner.bailer@joanneum.at

C. Gurrin · Ly-Duyen Tran
Dublin City University, Dublin, Ireland
E-mail: cathal.gurrin@dcu.ie, ly.tran2@mail.dcu.ie

B. Þ. Jónsson
IT University of Copenhagen, Copenhagen, Denmark
E-mail: bjth@itu.dk

J. Lokoč · F. Mejzlík · L. Peška · P. Veselý
Department of Software Engineering
Charles University, Prague, Czech Republic
E-mail: jakub.lokoc@matfyz.cuni.cz, frankmejzlik@gmail.com, ladislav.peska@matfyz.cuni.cz, prtrikvesely@gmail.com

A. Leibetseder · K. Schoeffmann
Klagenfurt University, Klagenfurt, Austria
E-mail: aleibets@itec.aau.at, ks@itec.aau.at

L. Rossetto
Departement of Informatics
University of Zurich, Zurich, Switzerland
E-mail: rossetto@ifi.uzh.ch

K. Schall
Visual Computing Group

Abstract The Video Browser Showdown addresses difficult video search challenges through an annual interactive evaluation campaign attracting research teams focusing on interactive video retrieval. The campaign aims to provide insights into the performance of participating interactive video retrieval systems, tested by selected search tasks on large video collections. For the first time in its ten year history, the Video Browser Showdown 2021 was organized in a fully remote setting and hosted a record number of sixteen scoring systems. In this paper, we describe the competition setting, tasks and results, and give an overview of state-of-the-art methods used by the competing systems. By looking at query result logs provided by ten systems, we analyze differences in retrieval model performances and browsing times before a correct submission. Through advances in data gathering methodology and tools, we provide a comprehensive analysis of ad-hoc video search tasks, discuss results, task design and methodological challenges. We highlight that almost all top performing systems utilize some sort of joint embedding for text-image retrieval and enable specification of temporal context in queries for known-item search. Whereas a combination of these techniques drive the currently top performing systems, we identify several future challenges for interactive video search engines and the Video Browser Showdown competition itself.

Keywords Interactive video retrieval, video browsing, video content analysis, content-based retrieval, evaluations

Acknowledgements This work was partially funded by the Czech Science Foundation (GAČR) under project 19-22071Y, by the EU's Horizon 2020 research and innovation programme under the grant agreements n^o 825079, MindSpaces, and n^o 951911, AI4Media - A European Excellence Centre for Media, Society and Democracy, by the FWF Austrian Science Fund under grant P 32010-N38, by the Swiss National Science Foundation (project "Participatory Knowledge Practices in Analog and Digital Image Archives", contract no. CRSII5_193788), and by Science Foundation Ireland under grant numbers 18/CRT/6223 and 18/CRT/6224.

1 Introduction

In the 21st century digital cameras decorate almost every corner in city centers and most pedestrians carry a smartphone capable of high quality video. While humankind has reached the point where digital video data are so easily produced, stored, and shared, a huge remaining challenge is effective and efficient access to

HTW Berlin, Berlin, Germany
E-mail: konstantin.schall@htw-berlin.de

L. Vadicamo
Institute of Information Science and Technologies (ISTI), CNR, Pisa, Italy
E-mail: lucia.vadicamo@isti.cnr.it

S. Vrochidis
Information Technologies Institute (ITI)
Centre for Research and Technology Hellas (CERTH), Thermi-Thessaloniki, Greece
E-mail: stefanos@iti.gr

J. Wu
Department of Computer Science
City University of Hong Kong, Hong Kong, China
E-mail: jiaxin.wu@my.cityu.edu.hk

Fig. 1: VBS2021 was organized as a fully virtual session.

these vast volumes of stored audio-visual information. So far, many commercial search engines have been established, allowing users to satisfy certain search needs over video collections with sufficient retrieval precision. Primarily, these search engines focus on returning matches to free-form text queries. However, high retrieval recall and interactive retrieval remain difficult challenges for current video search models.

The scientific community has reacted to the high recall challenge with evaluation campaigns attracting research teams focusing on video retrieval. TRECVID [37], Video Browser Showdown (VBS) [36], and Lifelog Search Challenge [13] define retrieval tasks where both high recall and precision are essential to achieve a good score. Every year, the results of these campaigns confirm that achieving high recall in arbitrary tasks over general videos remains a hard problem. So far, there was no clear solution to the problem, despite the limited scale of the competition datasets, compared to web-scale media collections. Nevertheless, one observation confirmed every year is that system-user interactions have a positive effect on effectiveness.

Two important task types for interactive retrieval evaluation are known-item search (KIS), where there is only a single correct item to be found, and ad-hoc video search (AVS), where the goal is to retrieve as many items as possible matching a description. This paper focuses on the Video Browser Showdown 2021, a virtual event (see Fig. 1) where a record number of participating teams tried to solve a large number of AVS and KIS tasks with their interactive video search systems. We emphasize that while user-centric evaluations of this kind and extent are rare and discrete events, they do provide invaluable insights to the performance of participating approaches. The key contributions of this paper can be summarized as:

- Description of VBS 2021, including an overview of participating systems and their rich set of tested approaches;
- Results of the first remote VBS 2021, where a record number of 16 scoring teams participated;
- Findings from the competition, comprehensive AVS task analysis, and result set log analysis;
- Critical analysis of current challenges with interactive AVS evaluations and suggestions for upcoming VBS evaluations.

The remainder of this paper is structured as follows: Section 2 gives an overview of VBS 2021 and its tasks, Section 3 introduces the participating systems and summarizes their approaches, Section 4 shows the results of the interactive evaluation with a particular focus on AVS analysis, and Section 5 gives an outlook towards the future and concludes the paper.

2 Video Browser Showdown

The Video Browser Showdown [36, 53], collocated with the International Conference on Multimedia Modeling (MMM), started its annual comparative live evaluations in 2012 and reached its tenth anniversary in 2021. Unlike other benchmark evaluations, VBS represents a unique evaluation platform where teams compete on a task at the same time, in the same environment, and with user-centric video search tools.

Like in previous years, VBS 2021 used the V3C1 [59] dataset, which contains approximately 1000 hours of video. The task types were unchanged, consisting of visual, where the target sequence was shown to participants, and textual, where the target sequence was described, known-item search (KIS) tasks and ad-hoc video search (AVS) tasks. For the sake of completeness, we will briefly recap the scoring function which was the same as in 2020 [36], albeit with minor adjustments. In KIS tasks, the goal is to reward quickly finding the correct item[1], while punishing wrong submissions. Given a linearly decreasing function f_{TS} based on search time, the time of correct submission t and the number of wrong submissions ws, the score for a given KIS task is as follows:

$$f_{KIS}(t, ws) = \lceil max(0, 50 + 50 \cdot f_{TS}(t) - 10 \cdot ws \rceil \tag{1}$$

f_{KIS} thus awards at least 50 points for a correct submission if no wrong submission was made, and penalizes each wrong submission with a malus of 10 points.

In AVS tasks, the goal is to reward both precision and recall. Given correct submissions C and incorrect submissions I of a team, all correct submissions of all teams for a task P and a quantization function q which merges temporally close correct shots into ranges,[2] the scoring function for AVS tasks is as follows:

$$f_{AVS}(C, I, P) = \left\lceil \frac{100 \cdot |C|}{|C| + \frac{|I|}{2}} \cdot \frac{|q(C)|}{|q(P)|} \right\rceil \tag{2}$$

[1] KIS tasks have a correct video sequence, a submission of any frame within the correct sequence counts as correct

[2] "since VBS 2018, ranges are fixed static non-overlapping segments of 180s duration" [61], in 2021 the ranges were dynamic and based on shot segmentation.

While the overall setting was very similar to previous events, VBS 2021 introduced two major novelties. First, and most importantly, the competition took place fully remotely due to the COVID-19 pandemic. This setting was facilitated by adopting the new 'Distributed Retrieval Evaluation Server' (DRES)[3] [54], which has been explicitly designed for such a distributed and scalable setup. Teams, consisting of two active participants each, could access the main screen of the server (displaying tasks and scores) via their browser, and submit results via a REST service to the central server instance. In addition, participants, judges, and organizers were connected in a video conferencing session for communication. The participants were also asked to provide a camera view that shows the screen of their VBS tool. The public VBS session was live-streamed on Twitch. Fig. 1 shows a screen-capture from the virtual event. While this setup relaxed the "same environment" setting, the teams nevertheless solved the tasks at the same time using the same dataset.

The second major novelty was a briefing session with the judges for AVS tasks before the competition, in which the task descriptions were discussed, and clarifications added. The aim was to eliminate ambiguities and ensure that the assessment of the judges is more consistent than in previous years. The task selection procedure was the same as described in [36]. This aim has not been fully reached, however, as some potential ambiguities become only apparent when seeing candidate results. Thus a trial-run involving stand-in participants might be useful for the judges to come to a common understanding of valid solutions to a task.

3 Participating Systems

Table 1 and Table 2 list the retrieval and interaction methods of the different systems at VBS 2021, respectively. In this section, we summarize the methods used and in doing so, also provide an extensive overview of state-of-the-art methods in multimedia retrieval. The categories used are similar to the ones from the 2020 review [36], with a new subsection added for interaction modalities, given that there were two virtual reality systems this year.

3.1 Text Search

The trend from previous iterations of VBS to textual queries [36, 53] continues this year. The effectiveness of embedding-based methods such as the W2VV++ model used by last year's winner, SOMHunter [26], as also shown in an evaluation of SOMHunter and vitrivr [52], makes such models a valuable addition to retrieval systems. The W2VV++ model and its variants [31, 34, 39] was integrated to all systems designed by the team from Charles University, namely VIRET, SOMHunter, VBS2020 Winner, and in the form of features for image search also to CollageHunter. VIRET used the CLIP model [45], VIREO the interpretable embeddings of the dual-task model [73], EOLAS a conventional textual embedding approach using autoencoders, VERGE an attention-based dual encoding model [12], and VISIONE the Transformer Encoder Reasoning Network (TERN) model

[3] DRES v0.8.1 was used, see: `https://github.com/dres-dev/DRES/releases/tag/0.8.1`

Table 1: Selected *search* approaches used by participating systems. For each system, a reference to the paper describing the method is given; V3C1 means metadata provided with the V3C1 dataset [59]. The ASR data for V3C1 was provided by [57]. Categories are similar to the 2020 VBS Analysis [36].

	overall score	solved KIS	shot detection	joint embedding	concepts	ASR	OCR	image search	sketch search	fusion of modalities	temporal query	relevance feedback
vitrivr [16]	**254**	**20**	V3C1	[67]	[57]	V3C1	[57]	[56]	[56]	[17]	[15, 17]	
VIRET [43]	**244**	**24**	[35, 66]	[34, 39, 45]				[34, 39]			[43]	
VIREO [74]	**235**	**21**	V3C1	[73]	[73]	V3C1	[65]	[73]	[41]	[73]	[74]	
SOMHunter [72]	**228**	**22**	[35, 66]	[34, 39]				[34, 39]			[35]	[9]
HTW [18]	**218**	**22**	[18]					[18]	[18]	[18]	[18]	[18]
CollageHunter [33]	**203**	**22**	[35, 66]	[34, 39]				[34, 39]			[35]	[9]
VERGE [4]	**183**	**19**	V3C1	[12]	[37, 76, 78, 69, 14, 23]			[44, 20]		[4]	[4]	
VBS2020 Winner [26]	**182**	**18**	[35, 66]	[34, 39]				[34, 39]			[35]	[9]
vitrivr-VR [67]	**179**	**17**	V3C1	[67]	[57]	V3C1	[57]	[56]		[17]		
Exquisitor [24]	**138**	**18**	V3C1		[75]	(V3C1)					[24]	[25]
VISIONE [2]	**106**	**16**	V3C1	[38]	[3, 46, 47, 77]			[38, 49]	[1, 46, 47, 77, 71, 5]	[1]	[2]	
diveXplore [29]	**93**	**11**	10s		[7, 68, 40, 78]			[30]				
VideoGraph [51]	**87**	**8**	V3C1		[57]	V3C1	[57]					
noshot [22]	**50**	**9**	1s		[46]							
IVIST [28]	**42**	**8**	V3C1		[8, 10, 14, 57]		[63]					
EOLAS [70]	**2**	**0**	V3C1			V3C1		[70]				

[38]. Both vitrivr and vitrivr-VR added a text co-embedding this year [67], based on an approach similar to W2VV++.

Concept-based search was also used by several teams this year. vitrivr, vitrivr-VR, and VideoGraph applied a combination of several neural networks [57] for concept detection. VideoGraph additionally contextualized and extended them by linking the extended concepts to Wikidata.[4] VERGE used a multitude of concept detection models, including EfficientNets trained on ImageNet1000 [11] and TRECVID SIN [37], EventNet [76], a style model [69] pre-trained models on MS COCO [32] and OpenImageV4 [27], a 3D-CNN model [14] pre-trained on the Kinetics-400 dataset [23], and VGG16 [64] trained on Places365 [78]. The last combination was also used by IVIST. Other concept detectors used include [3] by VISIONE, YOLO 9k [46] by NoShot, and EnlightenGan [21] combined with HTC [8] together with 3D ResNet-200 [14] by IVIST. VIREO utilized the decoded concept list of visual embedding [73]. HTW uses tagged image archives [18] to generate concepts, and Exquisitor uses pylucene to search the ResNeXt-101 visual concepts and their text descriptions [75] to provide positive examples to its relevance feedback process.

For ASR search, vitrivr, vitrivr-VR, VIREO, EOLAS, Exquisitor and VideoGraph all rely on the generated speech resource from the V3C1 dataset [59]. For OCR search, vitrivr, vitrivr-VR, and VideoGraph applied [57], VIREO tesseractOCR [65], and IVIST used ASTER [63].

[4] `https://www.wikidata.org`

3.2 Image and Sketch Search

For image similarity, VIRET, SOMHunter, VBS2020 Winner, and CollageHunter all used embedded W2VV++ model features [31,34,39]. VIREO uses visual embeddings of the dual-task model [73], VERGE the last pooling layer of a fine-tuned GoogleNet [44], HTW a CNN with DARAC-Pooling [60] and VISIONE Resnet101-GeM [49] and TERN [38]. For color or semantic sketches, vitrivr supports a plethora of features [50,55], VERGE clusters to twelve predefined colors using the Color Layout MPEG-7 descriptor, and HTW uses a handcrafted low-level feature [18]. VIREO [41] and VISIONE [1] also support sketch search, with VISIONE extracting dominant colors with pretrained color hash tables [5,71] and objects using pretrained neural networks [46,47,77]. CollageHunter allows image collages, which enable localization of example image queries on a canvas. In diveXplore, similar video summaries can be retrieved by image feature similarity [29]. EOLAS employs an image search mechanism using the positions of the user and the shots chosen in an embedded latent space, which is based on the assumption that shots that are similar are closer together. Exquisitor, VideoGraph, noshot, and IVIST do not include modules for image or sketch search.

3.3 Fusion Approaches

Multiple teams offer the option to formulate a query with a temporal modality. In vitrivr, users can specify multiple temporally ordered queries which are independently evaluated, and then aggregated with the scoring function rewarding videos which have matching segments for the individual queries in the correct order [17]. VIRET uses a context-aware ranking model [43] which requires that all independently formulated queries should be sufficiently answered by a segment of a video. Many teams allow users to specify two ordered queries, which are then executed independently. SOMHunter, VBS2020 Winner and CollageHunter all use the same algorithm as in 2020 [35], where the score for an item is determined by fusing its own score with the score of the best match for the second query within a specified time delta. HTW and VERGE used a similar algorithm for temporal queries. Similarly, for VISIONE, two independent queries describing two distinct keyframes of a target video can be submitted by the user; only the top 100,000 results for each query are retained and results from the same video which are within a specified time threshold are paired in the result visualization. The resulting pairs are ranked using a scoring function defined as a normalized sum of the scores of the outcomes in the pair. In VIREO, no temporal distance is specified, the ranking algorithm looks for sequences with the highest combined rank, ignoring temporal distance [42].

Besides the temporal context, there are also systems which offer different query modalities to the user. vitrivr and vitrivr-VR both score result items for each modality separately, and then offer a configurable choice of max- or average-pooling the score over the different modalities, with average-pooling being used in the competition. In VISIONE, all modalities are mapped to text, which allows the usage of Apache Lucene as a search backend. Each modality is a sub-query and the Lucene QueryRescorer combines their search results [1]. VIREO uses a linear function to fuse ranking lists of concept-based search and embedding-based search

[73], and VERGE provides the option to re-rank the results of a search modality, based on the results of any other modality. Exquisitor supports fusing the results of semantic classifiers. The most common operation is intersection of classifiers, where videos are returned if they have some keyframes ranked highly in both classifiers. Intersection can then be augmented by a temporal constraint, where a keyframe from one model must precede a keyframe from another by a specified minimum or maximum number of segments.

3.4 Relevance Feedback

While some teams offered support for simple more-like-this queries, such as vitrivr using deep features based on MobileNet V1[5] [19], there were also more sophisticated approaches to relevance feedback.

The goal of the Exquisitor system is to study the role of interactive learning in large-scale multimedia analytics applications. To that end, Exquisitor relies on user relevance feedback as its main user interaction strategy. The general goal of interactive learning is to develop a semantic classifier that captures the information need of the user well [25]. At the search-oriented VBS competition, however, the goal of this interaction is to build a classifier that can identify the most likely solution candidates, allowing the user to then explore the candidates in more detail to determine their relevance to the task.

SOMHunter, VBS2020 Winner and CollageHunter all use the same approach [9] as in 2020, which is a "Bayesian-like update rule to maintain current relevance scores of frames based on selected positive and implicit negative examples" [36].

3.5 Result Set Visualization and Browsing

Turning to the user interaction strategies presented in Table 2, the most common approach is still to present query results in an ordered list of small thumbnails representing keyframes (similar to previous iterations of VBS). The temporal context of results can then often be inspected based on user input, e.g., as a video preview or by browsing neighboring keyframes. This is also the approach used by the highest-scoring team, vitrivr. Some systems also offer a video player, or an option to view a summary of the entire video. Several teams have experimented with different browsing or visualization approaches.

Rather than displaying individual frames, VIRET focuses on displaying top-ranked video segments (i.e., fixed-length sequences of consecutive frames extracted from a video), where the best per-segment answers for each sub-query are visually highlighted. All three systems relying on the SOMHunter engine provide three result set visualization modes: ranked list of frames, ranked list of scenes (i.e., matched frame with its temporal context per row), and a self-organized map (SOM) evaluated dynamically over all scored database frames. The SOM-based display allows exploratory investigation of the result set, providing more diverse but semantically collocated items in the result set grid view.

[5] `https://tfhub.dev/google/imagenet/mobilenet_v1_050_192/quantops/feature_vector/3`

Table 2: Selected *interaction* approaches integrated and frequently used in the participating systems. A ✓ symbol indicates implementation in a given system. Categories are the same as in the 2020 Analysis [36].

	top-k from video filter	temporal context	video preview	video summary	video player	2D map embedding
vitrivr	✓	✓	✓	✓	✓	
VIRET	✓✓	✓✓		✓		
VIREO	✓✓	✓✓	✓		✓	
SOMHunter	✓✓	✓✓		✓		✓
HTW	✓✓	✓✓	✓	✓	✓	✓✓
CollageHunter	✓	✓✓		✓		✓
VERGE		✓✓	✓		✓	
VBS2020 Winner	✓	✓✓		✓		✓
vitrivr-VR	✓✓	✓✓	✓	✓	✓	
Exquisitor	✓	✓✓			✓✓	
VISIONE		✓✓		✓	✓✓	
diveXplore		✓✓	✓	✓✓	✓✓	
VideoGraph		✓	✓✓	✓✓	✓✓	
noshot			✓✓	✓	✓✓	
IVIST	✓		✓✓		✓✓	
EOLAS	✓		✓		✓	

HTW enables browsing of the whole video collection on keyframe- or shot-level by arranging the images on a hierarchical self-sorting map (SSM) [18]. Furthermore, the top-2,000 results are presented in either a list, a hierarchical SSM or video summary consisting of five shots.

The diveXplore system introduces a new way of browsing video summaries. Search results for this mode contain lists of videos appearing in panels that contain all shots as thumbnails. These panels can be browsed horizontally by search concept ranking and vertically by video summary similarity to the entire database.

The two VR systems used different approaches, which we discuss in the next subsection.

3.6 Interaction Modalities and Paradigms

The user interface of a retrieval system has a large impact on its performance by enabling and restricting interaction modalities. In this iteration of the VBS, for the first time, not all systems used a conventional desktop-based user interface, as EOLAS and vitrivr-VR became the first two systems to participate in VBS with virtual reality-based user interfaces.

Virtual reality as multimedia retrieval user interface offers both opportunities as well as challenges when compared to conventional desktop user interfaces. With the trend towards deep learning-assisted textual queries, VR interfaces require alternative text-entry methods in the absence of a physical keyboard. Both EOLAS and vitrivr-VR employ speech-to-text as the primary text entry method. vitrivr-VR additionally uses a direct interaction-based virtual keyboard as backup text entry method.

The approaches of EOLAS and vitrivr-VR differ the most to the other teams in regards to results visualization and interaction. EOLAS visualizes results as

Table 3: Overview of the scores for the individual task types per team (top-2 written in bold typeface).

Team	AVS	Visual KIS	Textual KIS	Overall Score
vitrivr	**100**	71	83	**254**
VIRET	50	**100**	94	**244**
VIREO	**80**	84	71	235
SOMHunter	44	83	**100**	228
HTW	36	82	**99**	218
CollageHunter	43	**85**	75	203
VERGE	34	70	80	183
VBS2020 Winner	39	70	73	182
vitrivr-VR	39	65	74	179
Exquisitor	23	58	58	138
VISIONE	20	65	21	106
diveXplore	20	39	34	93
VideoGraph	21	26	40	87
noshot	8	43	0	50
IVIST	13	29	0	42
EOLAS	2	0	0	2

clusters in 3D space, laid out according to their feature similarities, which can be traversed to explore the result set. vitrivr-VR employs a more conventional approach to result set visualization, by displaying the result set in a sorted grid, wrapped cylindrically around the user. In addition to a standard video player in VR, vitrivr-VR additionally makes use of virtual space by providing a video segment summary display resembling a file cabinet drawer, which allows quickly riffling through a temporally ordered box containing the segments of a video.

4 Results of VBS 2021

In this section, we present the results of the competition, and provide an analysis of submissions and retrieval models. Additionally, we are able to analyze AVS data for the first time since 2018, and provide insights into both system performance and task properties. The availability of AVS data is one of the reasons we focus on AVS tasks, KIS tasks are also analyzed in depth in previous papers [36, 53]. We exclude one participating system altogether [48], as the team experienced technical difficulties on both days of the evaluation. Analysis regarding result logs is only available for a subset of teams, since not all teams logged their results in the common format.

4.1 Overall Results

Table 3 shows an overview of all teams and the scores achieved per category, highlighting the top two scores per category. Scores are normalized per category such that the best team receives 100 points in said category. Categories are scored independently, the overall score is calculated by summing up the individual categories.

Table 4: Overview of the number of solved KIS tasks for the known-item search tasks per team (top-2 scores written in bold typeface). There were 21 V-KIS and 6 T-KIS tasks.

Team	Solved V-KIS tasks	Solved T-KIS tasks
vitrivr	16	**4**
VIRET	**20**	**4**
VIREO	18	3
SOMHunter	18	**4**
HTW	18	**4**
CollageHunter	**19**	3
VERGE	16	3
VBS2020 Winner	15	3
vitrivr-VR	14	3
Exquisitor	15	3
VISIONE	15	1
diveXplore	9	2
VideoGraph	6	2
noshot	9	0
IVIST	8	0
EOLAS	0	0

When looking at Table 3, the highest scoring team is different for every task category, and no team is among the two top-scoring systems of more than one category. This is an indication of well-designed tasks and meaningful differences between the top systems and their operators.

Comparing the scores of the two VR systems, EOLAS and vitrivr-VR shows that while VR can be competitive, the approach used by EOLAS for the user interface was not very suitable for the competition format. EOLAS's interface focused on exploring in a 3D environment involving VR locomotion, which caused difficulties in finding a sufficient number of shots in a limited time.

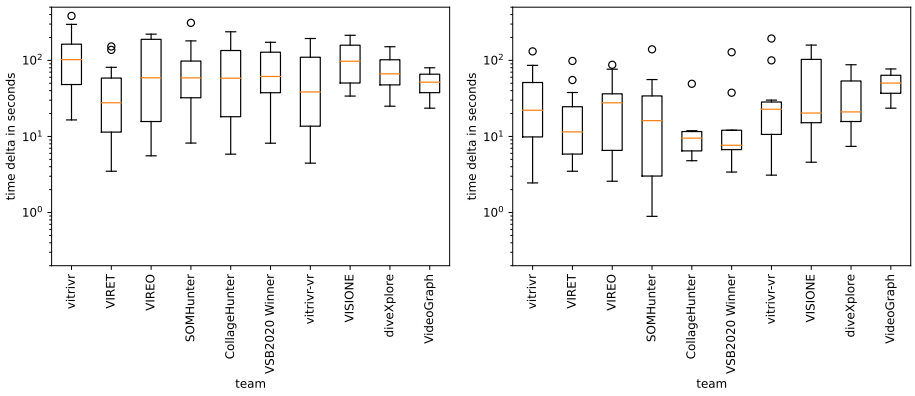
Most teams were able to solve a substantial number of Visual and Textual KIS tasks, as shown in Table 4. The easiest task was solved by 15 out of 16 of teams, and the most challenging one was not solved by any team. Across all tasks, the mean number of teams which solved a task was approximately 9.4.

4.2 Result Log Analysis

In addition to the submissions, most teams logged the result sets of their queries, either storing the logs locally or sending them directly to the competition server. In this section, we take a closer look at the logs, giving insight into the retrieval models and the differences in systems and operators.

4.2.1 Browsing Efficiency

One interesting question is how long it took operators to find an item once it was present in a result set. This is both dependent on the system, i.e., how good the browsing capabilities of a system are, and on the operator, since some operators



(a) First Appearance (b) Last Appearance

Fig. 2: Time deltas of teams between first and last appearance of correct item in result logs before submission on shot level.

prefer to browse a result set exhaustively, while others prefer to reformulate and execute new queries.

Figure 2 shows the elapsed time between the first and last appearance of the correct shot in the result set and submission time of the correct item. Note that it is possible that between one user receiving the correct result from their query and submitting it, the other user formulated a query which contained the correct result, and hence the time delta between last appearance and submission may not reflect the browsing time accurately. It is also possible that a correct item was found through the video and not the shot.

To visualize the dependency between the rank of a found item and the time until correct submission, we show in Figure 3 each correct submission as a datapoint with the rank it was found at first, and the time it took until correct submission. Overall, the figure shows that, as expected, the time between the first appearance and a correct submission increases. However, the figure also demonstrates that variance increases as well, indicating that operator differences are indeed occurring: while some operators might have browsed for a long time, others reformulated their query or found the correct item through the correct video.

We have conducted several other analyses, such as only considering items below a certain cutoff (which could be considered browsable), or considering the appearance of the best rank. These analyses have not produced new insights, and hence are omitted from the paper. The absence of standardized interaction logging which could indicate scrolling and currently visible results, makes this analysis challenging.

4.2.2 Comparison of Retrieval Models

For the comparison of retrieval models, Figure 4 shows where the best achieved rank of a correct item before submission was per system across tasks.

Figure 4 shows that the retrieval model strengths of the top teams are somewhat matched, with VIRET and SOMHunter finding the desired items in the first

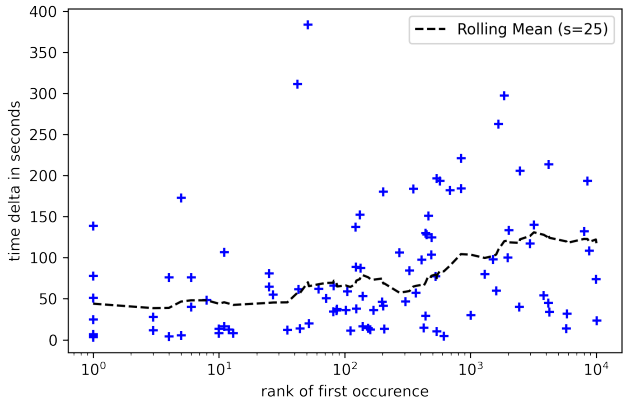


Fig. 3: Relation between the rank of first occurrence of a shot in the result logs and time delta to correct submission. As expected, time delta increases with rank, with variance increasing as well.

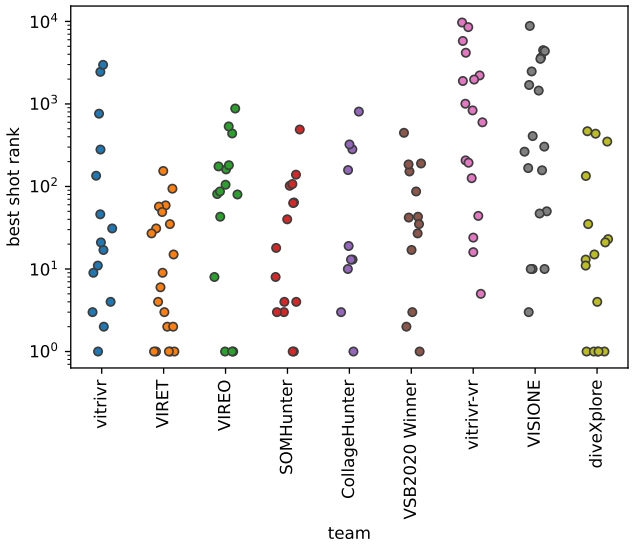


Fig. 4: Best rank of correct item appearing in result log. Teams are ordered by descending score on the x-axis. Different teams have used different thresholds to log results, for UX and performance reasons.

ten results more consistently. vitrivr, VIREO and CollageHunter have a lower sample size, which is explained by the fact that they were often able to find the correct item through a video-level hit and subsequent browsing (see Figure 5. Also of note is that in a previous evaluation of vitrivr and SOMHunter, results clearly showed that SOMHunter had a better retrieval model [52]. In the meanwhile, vitrivr added a joint embedding and improved its temporal scoring, allowing it to be competi-

Team	Metric	v-1	v-2	v-3	v-4	v-5	v-6	v-7	v-8	v-9	v-10	v-11	v-12	v-13	v-14	v-15	v-16	v-17	v-18	v-19	v-20	v-21	t-1	t-2	t-3	t-4	t-5	t-7
vitrivr	rs	2447	17	1	-	-	-	3	-	762	-	-	-	-	-	11	9	135	-	2978	-	4	46	2	31	280	21	-
	rv	1	17	1	62	50	34	3	8	6	191	30	48	2	23	11	9	135	20	7	97	1	46	2	26	72	21	12
	t	27	39	124	45	19	37	104	134	59	261	92	28	299	40	193	56	24	87	70	184	54	205	62	125	203	324	111
	tcs	67	145	134	75	102	102	-	144	117		175	40		47	210	78	-	130	187		107	403	72	287	334	-	-
VIRET	rs	1	3	59	-	-	94	6	4	1	-	1	1	35	-	-	2	9	27	49	1	154	15	2	-	31	-	57
	rv	1	3	56	16	4	4	6	4	1	37	1	1	35	1	3	2	9	9	6	1	1	1	2	1	7	16	1
	t	22	33	67	37	21	24	17	32	19	24	41	19	41	134	92	23	11	44	44	28	21	75	33	21	303	399	285
	tcs	100	44	166	134	28	54	25	36	24		48	23	61	141	128	30	25	99	94	35	35	97	56	153	416	-	-
VIREO	rs	-	-	81	-	-	-	1	161	80	181	880	439	-	-	87	105	175	-	534	-	1	43	-	-	1	-	8
	rv	26	26	81	9	39	39	1	10	12	101	63	1	1	13	5	51	2	1	3	67	1	1	35	13	1	-	8
	t	78	51	17	27	48	105	14	32	203	275	290	20	60	251	38	36	22	24	146	72	85	141	291	31	144	-	382
	tcs	104	58	-	40	55	141	19	99	247	-	-	49	70	268	41	96	30	40	261	95	87	227	-	91	255	-	-
SOMHunter	rs	102	-	18	-	-	-	64	4	3	107	139	-	3	-	-	1	1	-	-	63	-	40	4	-	490	8	-
	rv	9	39	16	72	11	2	19	4	1	107	1	1	3	2	1	1	1	13	1	4	2	16	4	2	4	8	70
	t	66	26	65	49	84	166	133	20	28	147	15	20	173	68	109	20	71	18	88	264	38	137	175	22	28	223	128
	tcs	102	140	90	77	89	229	166	62	30	-	68	47	177	-	126	21	78	34	138	-	53	322	195	65	153	-	-
CollageHunter	rs	-	-	13	-	-	-	13	-	-	-	808	-	-	-	-	3	323	-	-	158	-	-	19	-	1	282	10
	rv	1	5	13	14	3	9	5	4	1	55	17	1	6	9	2	3	35	6	1	1	1	1	19	9	1	19	1
	t	39	47	179	114	88	27	35	74	82	224	39	131	34	41	75	48	110	28	45	292	70	147	213	27	85	166	116
	tcs	119	73	190	213	95	90	44	164	115	252	44	146	-	53	84	53	122	58	82	-	113	177	262	38	-	-	-
VBS2020 Winner	rs	43	-	152	-	-	-	87	35	27	-	-	-	-	-	-	1	17	447	-	-	-	42	3	-	2	190	185
	rv	14	52	73	12	6	74	16	35	27	2	7	1	1	2	1	1	17	1	2	12	11	2	3	27	2	62	6
	t	138	253	281	63	13	13	38	37	113	127	14	26	221	93	70	15	130	145	77	15	37	36	31	62	47	263	72
	tcs	200	-	-	-	24		75	49	156		38	48	-	103	98	22	147	273	99	46	60	-	196	240	81	-	-
vitrivr-vr	rs	-	1006	126	5795	-	-	24	44	4172	-	-	-	-	8526	-	835	16	-	1975	-	5	599	1897	207	194	2215	9704
	rv	4	27	21	3	5	168	24	44	239	2547	50	1	20	14	12	9	16	6	395	1301	5	17	102	16	194	83	995
	t	80	35	56	31	33	38	33	28	32	105	162	32	252	84	229	34	68	26	54	31	31	55	67	69	183	97	197
	tcs	84	65	-	45	41		37	41	77		-	55	268	278		-	96	77	154		59	180	-	83	218	-	-
VISIONE	rs	1697	1452	50	167	-	-	3	47	10	3520	10	-	-	4370	-	408	10	-	304	-	8819	264	157	2478	4484	-	3643
	rv	11	28	1	1	29	61	1	1	10	61	1	1	53	5	3	29	10	3	2	111	5	48	157	10	31	63	4
	t	14	156	275	45	35	233	190	48	44	13	189	20	221	47	99	69	256	35	19	97	83	414	137	93	152	359	354
	tcs	142	-	-	69	45		205	53	64	-	236	33	230	107	128	166	-	114	65		192	-	-	298	-	-	-
diveXplore	rs	-	-	-	1	-	1	436	35	13	1	-	1	-	350	15	-	11	4	134	-	21	467	23	1	-	-	-
	rv	-	-	-	1	-	1	436	1	13	1	-	1	-	350	15	-	11	4	134	-	1	467	23	1	-	-	-
	t	-	-	-	111	-	12	33	179	65	133	-	104	-	49	286	-	83	14	66	-	100	403	293	243	-	-	-
	tcs	-	-	-	128	-	63	-	194	-	158	-	121	-	-	-	-	90	90	154	-	118	-	-	382	321	-	-
VideoGraph	rs	-	-	10087	-	-	-	1059	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1299	-	-	-	5447	-
	rv	-	-	12	8	-	33	1059	44	175	4405	-	-	-	-	-	-	-	-	-	-	-	91	6437	27	3898	2373	137
	t	-	-	117	86	-	37	42	68	47	169	-	-	-	-	-	-	-	-	-	-	-	239	154	284	79	97	99
	tcs	-	-	141	131	-	-	-	-	-	-	-	80	-	117	-	-	-	108	170	-	-	319	-	306	-	-	-

Fig. 5: Green cells show the best achieved logged rank r_s between 1 and 300 in time t of a correct scene frame in a task. The best rank r_v of a correct video frame from the same result log is included, while t_{cs} presents the time of the tool's correct submission. Red values are for the best detected ranks of searched video frames if searched scene frames were not present in the logged result sets for a task. Red or orange cells show a browsing failure where the frame or video was retrieved but the team did not submit a correct result.

tive again in the retrieval model and having to rely less on browsing. vitrivr-VR also used the joint embedding but lacked the ability to specify temporal context, which explains the lower ranks compared to vitrivr, even though both systems had access to the same features, indicating that having an easy way for users to specify temporal context in a query is essential for successful interactive video retrieval.

4.2.3 Analysis of Submissions

A more comprehensive overview of the result logs is shown in Figure 5, which shows best logged rank of the correct shot and video, the time it took for the item to appear at the given rank and the time of the correct submission. It also shows browsing misses, meaning the correct item (cell colored in red) or video (cell colored in orange) was present in the result set, but not submitted. Note that the logs for some teams, such as VideoGraph, can be incomplete due to technical difficulties.

The data shows that a substantial number of teams had video-level browsing misses, meaning the correct video was found, but not the correct segment. Shot-

level misses were rarer, but still happened, e.g. for vitrivr and SOMHunter in three tasks, with the rank of the correct shot ranging from 1 (t-7, CollageHunter) to 9704 (t-7, vitrivr-VR). While missing the correct item at rank 1 is a browsing-level miss which can be attributed to the operator (and also the result visualization component), when missing an item at higher ranks it is not knowable, with the current logging specification, whether the operator browsed that far or whether they simply formulated another query after looking at a subset of high-ranked results.

Additionally, many correct submissions originated from a video-level hit, with operators subsequently exploring the video through neighboring frames, a video overview or with a video player. These cases are indicated by red numbers in Figure 5 and show that the ability to inspect a video is key to good performance in KIS tasks.

4.3 AVS Analysis

In the 2019 and 2020 iterations of VBS, there was no analysis of AVS tasks due to technical issues [36,62]. This year, the new evaluation server [54] improved testing by teams before the competition, which helped improve data quality. In this section, we are therefore able to present insight into questions surrounding AVS tasks.

In addition to retrieval, the judgement of AVS submissions is also done interactively at VBS. This has so far in every year resulted in different understandings, both between different judges and between judges and teams. Additionally, some tools, such as VISIONE, had issues with result submission, partially due to network overload and partially due to suboptimal implementations. We believe these issues did not significantly affect the results discussed in this section, which are presented in an aggregate form, as the number of submissions that were not submitted successfully by the affected teams is only a small fraction of the total number of submissions made by all teams.

Table 5 shows all AVS tasks and their description in the order which they were solved in the competition. All plots going forward include the task identifiers.

4.3.1 Judgement and Submissions During Tasks

One area of interest is how the assessed correctness of submissions changes during the time allocated to a task. The hypothesis being that at the start of a task, there is some ambiguity between the task description and judge and operator understanding of the description, which is resolved as teams see thumbnails of submissions judged as correct or incorrect.

In Figure 6, we show the ratio of submissions judged as correct over time. What stands out is that there were two tasks with a large degree of difference in task understanding, a-3 (person skiing with their own skis in the picture) and a-11 (person skiing, camera looking into the sun). For a-3, the difference (the task intention was for point-of-view shots) was clarified with a comment from a judge, however the ratio remains low since not all teams followed the discussion. For a-11 the different understandings persisted. Overall, no clear trend emerges. Some tasks

Table 5: List of all AVS tasks with their description, ordered by appearance order in the competition (a-5 was solved first, a-6 last)

Task ID	Task Description
a-5	Find shots of a person holding or waving a flag.
a-9	Find shots of at least one person drinking beer.
a-8	Find shots inside an airplane, showing at least one passenger.
a-1	Find outdoor shots of two women walking and talking to each other.
a-2	Find shots of people having their hair done.
a-3	Find shots of a person skiing, with his/her own skis in the picture.
a-10	Find shots of two adult men hugging each other.
a-4	Find shots of kids playing football (soccer).
a-11	Find shots of people skiing, shot with the camera looking into the sun (back-lit shot, possibly with lens flare).
a-6	Find underwater shots of one or more fish.

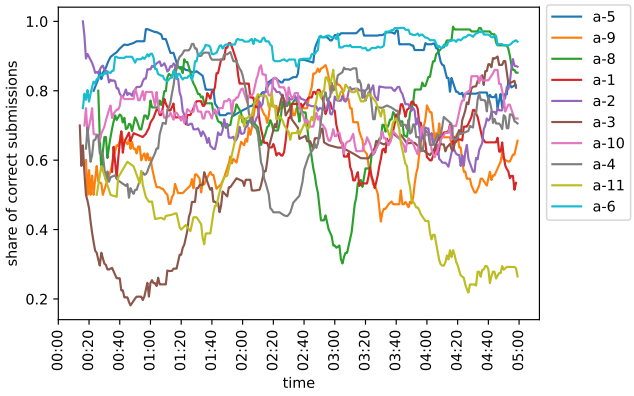


Fig. 6: Share of AVS submissions judged as correct over time during an AVS task.

exhibit consistently high agreement (e.g. a-6, looking for fish underwater and a-5, person with a flag), while most tasks have a high variance during the task.

Since the previous plot does not indicate the quantity of submissions, in Figure 7, we show how the number of submissions varies over time. Looking at the figure, it seems that some time is needed until a query is found which is suitable for the task at hand, and afterwards the rate of submissions stays relatively steady over time. This poses the question at which point in time there would be a drop-off in the number of submissions, if the AVS tasks had a longer duration.

Figure 8 shows that in addition to the rate of submissions remaining steady, the number of unique correct videos that are found also continues to increase

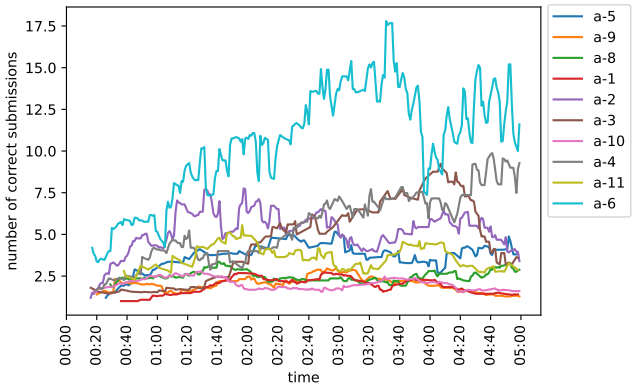


Fig. 7: AVS Submissions over time with a mean sliding window of size 25.

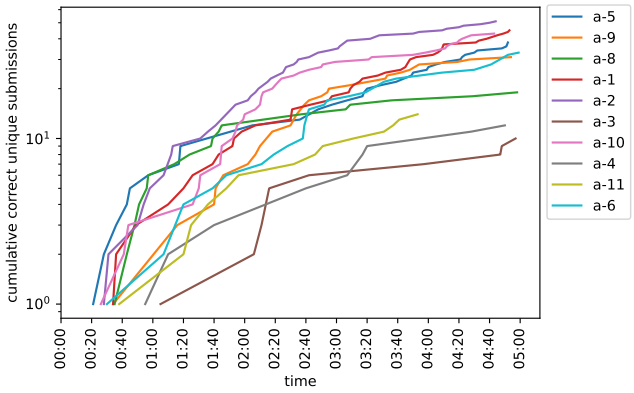


Fig. 8: Cumulative unique correct video submissions over time during an AVS task.

towards the end of the task, showing that even at the end of the time limit, new videos matching the description are still being found. This indicates that given a longer task duration, the number of unique correct submissions would probably still increase, as long as there exist relevant segments in the collection.

4.3.2 Differences Between AVS Tasks

Another interesting question is what differences, if any, there are between AVS tasks. For some tasks, looking at a thumbnail is sufficient (e.g., underwater shot of fish), while for tasks describing an action, the video needs to be inspected (e.g., shots of two women walking and talking). Additionally, some tasks might have a very wide range of acceptable results, while others are quite narrow in their description.

Figure 9, which focuses on all submissions, and Figure 10, which focuses on correct submissions, show the difference between the AVS tasks in terms of selected metrics: the number of overall submissions (shown as bars), time until first

(correct) submission, time to first (correct) submission by half the teams, and time until first ten (correct) submissions by half the teams. The y-axis indicating the time, on the right, has been inverted, so that higher y-axis values indicate that a task is easier for all metrics. On the x-axis, tasks are ordered by their appearance in the competition, with a-5 being the first task solved.

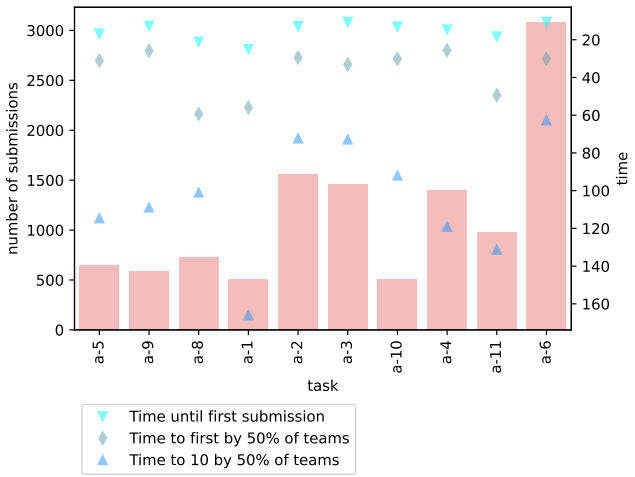


Fig. 9: Selected AVS metrics per task. Higher y-axis values indicate that for a task, teams found it easier to find results to submit.

Looking at these three graphs, the data indicates that there are relevant differences between the AVS tasks. For example, looking at a-1, it took almost five minutes for half of the teams to find 10 submissions which were judged as correct.

Additionally, we are interested in which kind of strategies are rewarded by the current evaluation metrics. In Figure 11, we show the performance of each team per task as a colored dot, with the color indicating the score in that task. The figure shows that the current scoring scheme seems to reward recall, in that teams which have a high share of overall submissions get higher score, even at lower precision.

4.3.3 Judge and Team Agreement Analysis

In Table 6, we show numbers of submissions where several teams agreed or disagreed with a judgement for a shot. Each column represents the number of teams with the same opinion about a particular submission. The first column shows that there are many unique submissions by teams and that there are frequent one-to-one disagreements. Although teams might prefer risky submissions, there might be also uncertain cases depending on text interpretation. The second column shows that in 80% of tasks two teams agree with a judge more often than two teams disagree with a judge.

Looking at extreme cases, in tasks a-2 and a-10 there was a correct submission provided even by eleven teams. This indicates that there might be a clear match

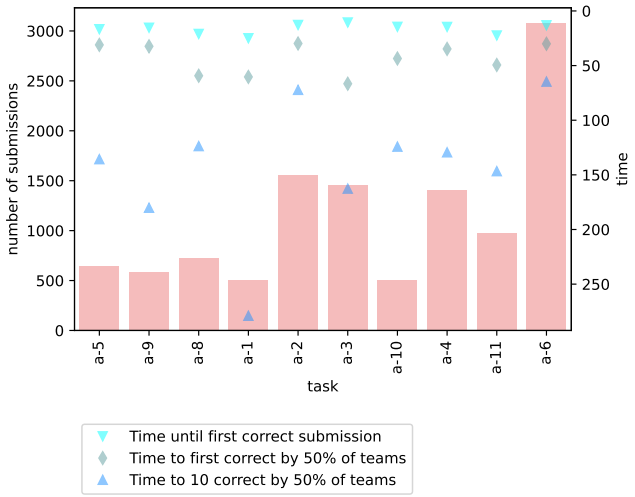


Fig. 10: Selected AVS metrics per task, looking at correct submissions. Higher y-axis values indicate that for a given task, it is easier to find results which judges deem correct.

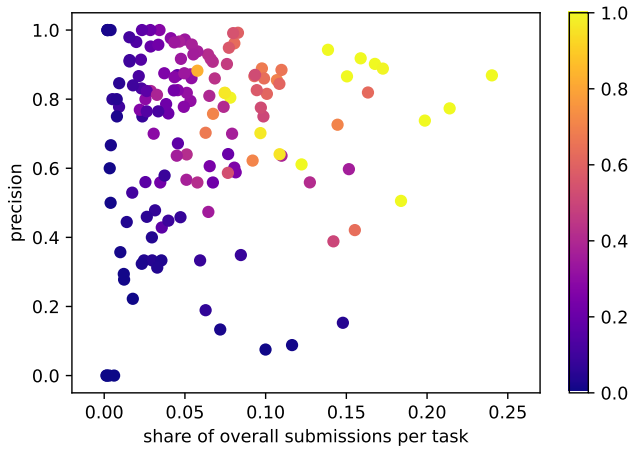


Fig. 11: Share of overall submissions per task over precision per team and task with the color indicating the evaluation metric score normalized over the best score of a task. For each task, all teams are represented as a dot.

between an AVS task text description and the visual content of an easy-to-find shot. This high level of agreement can be observed among multiple other tasks. On the contrary, the task a-3 represents an example of frequent disagreement between 2 to 4 teams and the decision provided by a judge. Indeed, many teams misunderstood the task a-3 asking to "Find shots of a person skiing, with his/her own skis in the picture" and did not realize the first-person view is required. In

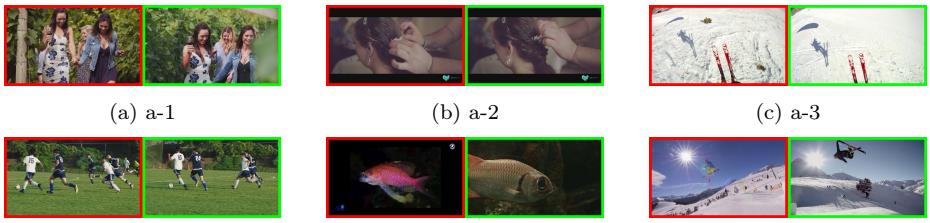
order to prevent this issue in future evaluations, there are two options: Either such a task is better defined (e.g., example image, discussion) or the task is excluded from scoring after the task is performed. In order to indicate a problem with an AVS task, statistics present in Table 6 could also be automatically reported by the evaluation server.

Table 6: The number of distinct correct / incorrect submissions where 1 - 11 teams agreed / disagreed with judges. Bold font highlights cases where the fraction is lower or equal to one (i.e., $\frac{\#agreement}{\#disagreement} \leq 1$).

	Number of teams in agreement / disagreement with judges										
task	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
a-1	**66 / 126**	9 / 7	4 / 1	2 /	-	1 /	2 /	-	-	-	-
a-2	340 / 226	108 / 51	60 / 13	18 / 2	15 /	10 / 1	6 /	5 /	5 /	3 /	1 /
a-3	**342 / 476**	**64 / 89**	**17 / 17**	**6 / 6**	-	-	-	-	-	-	-
a-4	**84 / 184**	**24 / 26**	10 / 4	9 / 1	3 /	2 /	-	-	-	-	-
a-5	122 / 41	30 / 3	20 / 1	13 / 1	2 / 1	2 /	2 /	1 /	1 /	-	-
a-6	863 / 228	336 / 8	188 / 1	79 /	35 /	15 /	6 /	2 /	-	-	-
a-8	102 / 85	39 / 16	18 / 2	13 / 1	8 /	2 /	3 /	1 /	-	-	-
a-9	**70 / 125**	39 / 14	16 / 5	10 / 1	5 / 2	4 /	3 /	2 /	-	-	-
a-10	**79 / 96**	40 / 2	20 / 1	13 / 1	5 /	4 /	1 /	1 /	1 /	-	1 /
a-11	**226 / 328**	55 / 49	32 / 2	12 / 2	7 /	1 /	1 /	-	-	-	-

4.3.4 Submission Similarity Analysis

After having analyzed the judge-team agreement, this section further investigates the inter-judge agreement through determining similar images that are judged differently. Figure 12 shows a selection of keyframe pairs exhibiting high similarities to each other, while judges disagree on their correctness. The similarities are determined by computing the Euclidean distance of the last fully connected layer vectors using Inception Net v3 [68]. By analyzing the distances per task, we find that the submissions for tasks a-1, a-5, and a-9 are less similar to each other compared to the other tasks, yet, all tasks contain questionable judgements. Similarly to above findings, task a-3 is interpreted very diversely, i.e. sometimes only a first-person view of ski tips are accepted and other times also third-person views of skiers on a slope. Such disagreements can be observed for other tasks as well; oftentimes different judgements are given on scenes that merely are a few shots apart (cf. Figures 12a-12d). In other cases, the submitted scenes are not related but, nevertheless, the judges' agreement on content correctness diverges (cf. Figures 12e-12f). Overall, when considering the lower 20% of all differently judged image distances per task, we identify an average of 109 similar items (excluding the outlier task a-3, which has 3,508 such items). Although not all similar yet differently scored images necessarily include misjudgements, it appears that the pre-task judge briefing done "on paper" was not an effective way to avoid them. Some differences in task understanding seem to become only apparent when seeing actual examples arriving. Thus a trial-run with judges or using multiple judgements with voting could be better alternatives.



(a) a-1 (b) a-2 (c) a-3

(d) a-4 (e) a-6 (f) a-11

Fig. 12: Similar but differently judged AVS submissions, judged as incorrect (red border, on the left) vs. judged as correct (green border, on the right).

5 Conclusion and Outlook

Ten years after the first Video Browser Showdown, the recent iteration had the largest number of participating systems so far. Despite organizational challenges with the fully remote setting, this iteration was very successful. For future evaluation campaigns, we see multiple challenges, which we will outline here.

On-Site versus remote VBS event. While the remote setting has many advantages, such as lower barrier to participation and cost, the conference setting had multiple key advantages which cannot be fully replicated in the remote setting, such as the ability of VBS to double as an informal demo session, where participants can try out other systems and ask questions. Networking and collaboration effects were diminished in comparison to previous years. Also, the novice session—which is a unique and important part of VBS—could not take place in the virtual setting.

Barrier to participation. While the barrier to participation in VBS remains somewhat high, this year had the highest number of participants so far, indicating that efforts to lower the barrier helped. Several pre-extracted features are also available for V3C1 [6,57] and V3C2 [58], enabling teams to focus on particular aspects. For completely new participants, it might be beneficial to further encourage authors to open source their systems. Currently, vitrivr is fully open-source,[6] and SOMHunter has an open-source release,[7] but full reproducibility of the competition would require all used systems to be open-source.

Result robustness. As demonstrated in [53], the difference in performance between users of the same system is rather large and increasing the number of users per system makes the results more statistically significant. In the current VBS format, two users operate the same system as a team. A larger number of users which solve tasks independently would increase confidence in the evaluation results and enable interesting analysis questions. This would however make it more difficult to consider systems which use explicitly collaborative retrieval strategies.

While this was the fifth year in a row that included AVS tasks, this time they have caused extended discussions amongst teams and organizers. Although the

[6] `https://vitrivr.org`

[7] `https://github.com/siret/somhunter`

queries for these tasks have been carefully selected and judges briefed in advance, we still encountered several difficulties for tasks with a high number of potentially correct results that could be discussed in a dedicated paper and should only be briefly mentioned here:

Disagreement: As Table 6 has shown, there is a substantial number of submissions where multiple teams perceived a segment as correct, but the judge disagreed. In Section 4.3.4 we also showed that semantically identical shots are sometimes judged differently by different judges, further underlining the challenge of AVS task evaluation. In future iterations of VBS we might consider a voting scheme to rectify this issue.

Evaluation Limits: Both the total number of submissions seen in Figure 7 and the number of unique correct videos seen in Figure 8 indicate that with a longer task duration, more items could be found. Due to the large number of submissions per task, however, extending the time for a single task would require more judges.

Resource Limits: While in previous years the old server software itself caused delays in submission processing, this year we faced severe network issues. Due to the fully virtual session and the high number of judges and participants—who were not only submitting many results, but also following the competition status via the server's web interface—the LAN and WAN limits (10 Gbps) of the server's location (Klagenfurt University) were reached. This unfortunately resulted in laggy behavior with packet losses and re-transmissions, slowing down the entire submission process.

Synchronous Submissions: In addition to the problems with the network load, some teams implemented their system such that submissions had to be confirmed by the server, which seriously limited their submission capacity due to the high network delay.

VBS 2021 successfully demonstrated that a fully virtual setting is feasible. In particular, for KIS tasks, the evaluation procedure went smoothly and almost all competing teams were able to solve some tasks, with most teams being able to solve more than 50% of KIS tasks.

There is still a large difference between the performance of the top teams, indicating no need to extensively modify task difficulty. With the move towards a larger dataset next year, we expect strong retrieval models to become more important, as approaches which rely on browsing must deal with twice as much data. At the same time, the evaluation procedure itself will become more challenging too, especially for AVS tasks which might have even more results (and need more judges) due to the larger dataset.

Even though VBS has been running for 10 years already, interactive video retrieval remains a hot topic with many challenges, which cannot be easily solved with only improved deep learning models. A strong focus on the retrieval efficiency, as well as the user interface, will be key to further push large-scale interactive video search.

6 Declarations

6.1 Funding

Funding information is provided in the acknowledgements section.

6.2 Conflicts of interest / Competing interests

Not applicable

6.3 Availability of data and material

The raw data used for the analyses presented in this paper are available via `https://zenodo.org/record/5566853`.

6.4 Code availability

Where evaluated system are open source, links are given in the paper references from Table 1. As mentioned, the source code for the evaluation server is provided on Github.[8] The full analysis code used to produce all graphs and tables is also available on Github.[9]

References

1. Amato, G., Bolettieri, P., Carrara, F., Debole, F., Falchi, F., Gennaro, C., Vadicamo, L., Vairo, C.: The visione video search system: Exploiting off-the-shelf text search engines for large-scale video retrieval. Journal of Imaging **7**(5) (2021). URL `https://doi.org/10.3390/jimaging7050076`
2. Amato, G., Bolettieri, P., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: VISIONE at video browser showdown 2021. In: International Conference on Multimedia Modeling, pp. 473–478. Springer (2021). URL `https://doi.org/10.1007/978-3-030-67835-7_47`
3. Amato, G., Falchi, F., Gennaro, C., Rabitti, F.: Searching and annotating 100M images with yfcc100m-hnfc6 and mi-file. In: Workshop on Content-Based Multimedia Indexing, pp. 26:1–26:4. ACM (2017). URL `https://doi.org/10.1145/3095713.3095740`
4. Andreadis, S., Moumtzidou, A., Gkountakos, K., Pantelidis, N., Apostolidis, K., Galanopoulos, D., Gialampoukidis, I., Vrochidis, S., Mezaris, V., Kompatsiaris, I.: VERGE in vbs 2021. In: International Conference on Multimedia Modeling, pp. 398–404. Springer (2021). URL `https://doi.org/10.1007/978-3-030-67835-7_35`
5. Benavente, R., Vanrell, M., Baldrich, R.: Parametric fuzzy sets for automatic color naming. JOSA A **25**(10), 2582–2593 (2008). URL `https://doi.org/10.1364/JOSAA.25.002582`
6. Berns, F., Rossetto, L., Schoeffmann, K., Beecks, C., Awad, G.: V3C1 dataset: An evaluation of content characteristics. In: International Conference on Multimedia Retrieval, pp. 334–338. ACM (2019). URL `https://doi.org/10.1145/3323873.3325051`
7. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. CoRR **abs/2004.10934** (2020). URL `https://arxiv.org/abs/2004.10934`

[8] `https://github.com/dres-dev/DRES`

[9] `https://github.com/vGsteiger/VBS-Analysis`

8. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: Hybrid task cascade for instance segmentation. Conference on Computer Vision and Pattern Recognition pp. 4969–4978 (2019). URL https://doi.org/10.1109/CVPR.2019.00511
9. Cox, I., Miller, M., Omohundro, S., Yianilos, P.: Pichunter: Bayesian relevance feedback for image retrieval. In: International Conference on Pattern Recognition, vol. 3, pp. 361–369. IEEE (1996). URL https://doi.org/10.1109/ICPR.1996.546971
10. Deng, D., Liu, H., Li, X., Cai, D.: Pixellink: Detecting scene text via instance segmentation. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), pp. 6773–6780. AAAI (2018)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009). URL https://doi.org/10.1109/CVPR.2009.5206848
12. Galanopoulos, D., Mezaris, V.: Attention mechanisms, signal encodings and fusion strategies for improved ad-hoc video search with dual encoding networks. In: International Conference on Multimedia Retrieval, pp. 336–340. ACM (2020). URL https://doi.org/10.1145/3372278.3390737
13. Gurrin, C., Jónsson, B.Þ., Schöffmann, K., Dang-Nguyen, D., Lokoc, J., Tran, M., Hürst, W., Rossetto, L., Healy, G.: Introduction to the fourth annual lifelog search challenge, lsc'21. In: International Conference on Multimedia Retrieval, pp. 690–691. ACM (2021). URL https://doi.org/10.1145/3460426.3470945
14. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? (2018). URL https://doi.org/10.1109/CVPR.2018.00685
15. Heller, S., Amiri Parian, M., Gasser, R., Sauter, L., Schuldt, H.: Interactive lifelog retrieval with vitrivr. In: Proceedings of the Third Annual Workshop on Lifelog Search Challenge, pp. 1–6 (2020). URL https://doi.org/10.1145/3379172.3391715
16. Heller, S., Gasser, R., Illi, C., Pasquinelli, M., Sauter, L., Spiess, F., Schuldt, H.: Towards explainable interactive multi-modal video retrieval with vitrivr. In: International Conference on Multimedia Modeling, pp. 435–440. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_41
17. Heller, S., Sauter, L., Schuldt, H., Rossetto, L.: Multi-stage queries and temporal scoring in vitrivr. In: International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–5. IEEE (2020). URL https://doi.org/10.1109/ICMEW46912.2020.9105954
18. Hezel, N., Schall, K., Jung, K., Barthel, K.U.: Video search with sub-image keyword transfer using existing image archives. In: International Conference on Multimedia Modeling, pp. 484–489. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_49
19. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR **abs/1704.04861** (2017). URL http://arxiv.org/abs/1704.04861
20. Jégou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(1), 117–128 (2010). URL https://doi.org/10.1109/TPAMI.2010.57
21. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing **30**, 2340–2349 (2021). URL https://doi.org/10.1109/TIP.2021.3051462
22. Karisch, C., Leibetseder, A., Schoeffmann, K.: Noshot video browser at vbs2021. In: International Conference on Multimedia Modeling, pp. 405–409. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_36
23. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. CoRR **abs/1705.06950** (2017). URL http://arxiv.org/abs/1705.06950
24. Khan, O.S., Jónsson, B.Þ., Larsen, M., Poulsen, L., Koelma, D.C., Rudinac, S., Worring, M., Zahálka, J.: Exquisitor at the Video Browser Showdown 2021: Relationships between semantic classifiers. In: International Conference on Multimedia Modeling, pp. 410–416. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_37
25. Khan, O.S., Jónsson, B.Þ., Rudinac, S., Zahálka, J., Ragnarsdóttir, H., Þorleiksdóttir, Þ., Guðmundsson, G.Þ., Amsaleg, L., Worring, M.: Interactive learning for multimedia at large. In: Proceedings of the European Conference on Information Retrieval. Springer (2020). URL https://doi.org/10.1007/978-3-030-45439-5_33

26. Kratochvíl, M., Veselý, P., Mejzlík, F., Lokoč, J.: Som-hunter: Video browsing with relevance-to-som feedback loop. In: International Conference on Multimedia Modeling, pp. 790–795. Springer (2020). URL https://doi.org/10.1007/978-3-030-37734-2_71
27. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J.R.R., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., Ferrari, V.: The open images dataset V4. International Journal of Computer Vision **128**(7), 1956–1981 (2018). URL https://doi.org/10.1007/s11263-020-01316-z
28. Lee, Y., Choi, H., Park, S., Ro, Y.M.: IVIST: Interactive video search tool in VBS 2021. In: International Conference on Multimedia Modeling, pp. 423–428. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_39
29. Leibetseder, A., Schoeffmann, K.: Less is more - divexplore 5.0 at VBS 2021. In: International Conference on Multimedia Modeling, pp. 455–460. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_44
30. Leibetseder, A., Schoeffmann, K.: lifexplore at the lifelog search challenge 2021. In: Proceedings of the Fourth Annual Workshop on Lifelog Search Challenge, pp. 23–28. ACM (2021). URL https://doi.org/10.1145/3463948.3469060
31. Li, X., Xu, C., Yang, G., Chen, Z., Dong, J.: W2VV++: fully deep learning for ad-hoc video search. In: International Conference on Multimedia, pp. 1786–1794. ACM (2019). URL https://doi.org/10.1145/3343031.3350906
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Computer Vision – ECCV, pp. 740–755. Springer (2014). URL https://doi.org/10.1007/978-3-319-10602-1_48
33. Lokoč, J., Bátoryová, J., Smrž, D., Dobranský, M.: Video search with collage queries. In: International Conference on Multimedia Modeling, pp. 429–434. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_40
34. Lokoč, J., Souček, T., Veselý, P., Mejzlík, F., Ji, J., Xu, C., Li, X.: A W2VV++ case study with automated and interactive text-to-video retrieval. In: International Conference on Multimedia. ACM (2020). URL https://doi.org/10.1145/3394171.3414002
35. Lokoč, J., Kovalčík, G., Souček, T., Moravec, J., Čech, P.: A framework for effective known-item search in video. In: International Conference on Multimedia, pp. 1777–1785. ACM (2019). URL https://doi.org/10.1145/3343031.3351046
36. 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 Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**(3) (2021). URL https://doi.org/10.1145/3445031
37. Markatopoulou, F., Moumtzidou, A., Galanopoulos, D., Avgerinakis, K., Andreadis, S., Gialampoukidis, I., Tachos, S., Vrochidis, S., Mezaris, V., Kompatsiaris, I., Patras, I.: ITI-CERTH participation in TRECVID 2017. In: TREC Video Retrieval Evaluation. NIST (2017). URL https://doi.org/10.5281/zenodo.1183440
38. Messina, N., Falchi, F., Esuli, A., Amato, G.: Transformer reasoning network for image-text matching and retrieval. In: International Conference on Pattern Recognition. IEEE (2020). URL https://doi.org/10.1109/ICPR48806.2021.9413172
39. Mettes, P., Koelma, D.C., Snoek, C.G.M.: Shuffled imagenet banks for video event detection and search. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **16**(2), 44:1–44:21 (2020). URL https://doi.org/10.1145/3377875
40. Monfort, M., Vondrick, C., Oliva, A., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L.M., Fan, Q., Gutfreund, D.: Moments in time dataset: One million videos for event understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 502–508 (2020). URL https://doi.org/10.1109/TPAMI.2019.2901464
41. Nguyen, P.A., Lu, Y.J., Zhang, H., Ngo, C.W.: Enhanced VIREO KIS at VBS 2018. In: International Conference on Multimedia Modeling, pp. 407–412. Springer (2018). URL https://doi.org/10.1007/978-3-319-73600-6_42
42. Nguyen, P.A., Wu, J., Ngo, C.W., Francis, D., Huet, B.: VIREO @ video browser showdown 2020. In: International Conference on Multimedia Modeling, pp. 772–777. Springer (2020). URL https://doi.org/10.1007/978-3-030-37734-2_68
43. Peška, L., Kovalčík, G., Souček, T., Škrhák, V., Lokoč, J.: W2VV++ BERT model at VBS 2021. In: International Conference on Multimedia Modeling, pp. 467–472. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_46
44. Pittaras, N., Markatopoulou, F., Mezaris, V., Patras, I.: Comparison of fine-tuning and extension strategies for deep convolutional neural networks. In: International Conference

on Multimedia Modeling, pp. 102–114. Springer (2017). URL http://doi.org/10.1007/978-3-319-51811-4_9
45. 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. CoRR **abs/2103.00020** (2021). URL https://arxiv.org/abs/2103.00020
46. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Conference on Computer Vision and Pattern Recognition, pp. 7263–7271. IEEE (2017). URL https://doi.org/10.1109/CVPR.2017.690
47. Redmon, J., Farhadi, A.: YOLOv3 on the Open Images dataset. https://pjreddie.com/darknet/yolo/ (2018). [Online; accessed 22-April-2021]
48. Ressmann, A., Schoeffmann, K.: Ivos-the itec interactive video object search system at vbs2021. In: International Conference on Multimedia Modeling, pp. 479–483. Springer (2021)
49. Revaud, J., Almazan, J., Rezende, R., de Souza, C.: Learning with average precision: Training image retrieval with a listwise loss. In: International Conference on Computer Vision, pp. 5106–5115. IEEE (2019). URL https://doi.org/10.1109/ICCV.2019.00521
50. Rossetto, L.: Multi-modal video retrieval. Ph.D. thesis, University of Basel (2018). URL https://doi.org/10.5451/unibas-006859522
51. Rossetto, L., Baumgartner, M., Ashena, N., Ruosch, F., Pernisch, R., Heitz, L., Bernstein, A.: Videograph - towards using knowledge graphs for interactive video retrieval. In: International Conference on Multimedia Modeling, pp. 417–422. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_38
52. Rossetto, L., Gasser, R., Heller, S., Parian-Scherb, M., Sauter, L., Spiess, F., Schuldt, H., Peska, L., Soucek, T., Kratochvil, M., et al.: On the user-centric comparative remote evaluation of interactive video search systems. IEEE MultiMedia (2021). URL https://doi.org/10.1109/MMUL.2021.3066779
53. Rossetto, L., Gasser, R., Lokoč, J., Bailer, W., Schoeffmann, K., Muenzer, B., Souček, T., Nguyen, P.A., Bolettieri, P., Leibetseder, A., et al.: Interactive video retrieval in the age of deep learning–detailed evaluation of VBS 2019. IEEE Transactions on Multimedia **23**, 243–256 (2020). URL https://doi.org/10.1109/TMM.2020.2980944
54. Rossetto, L., Gasser, R., Sauter, L., Bernstein, A., Schuldt, H.: A system for interactive multimedia retrieval evaluations. In: International Conference on Multimedia Modeling. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_33
55. Rossetto, L., Gasser, R., Schuldt, H.: Query by semantic sketch. CoRR **abs/1909.12526** (2019). URL http://arxiv.org/abs/1909.12526
56. Rossetto, L., Giangreco, I., Schuldt, H.: Cineast: A multi-feature sketch-based video retrieval engine. In: International Symposium on Multimedia, pp. 18–23. IEEE (2014). URL https://doi.org/10.1109/ISM.2014.38
57. Rossetto, L., Parian, M.A., Gasser, R., Giangreco, I., Heller, S., Schuldt, H.: Deep learning-based concept detection in vitrivr. In: International Conference on Multimedia Modeling, pp. 616–621. Springer (2019). URL https://doi.org/10.1007/978-3-030-05716-9_55
58. Rossetto, L., Schoeffmann, K., Bernstein, A.: Insights on the V3C2 dataset. CoRR **abs/2105.01475** (2021). URL https://arxiv.org/abs/2105.01475
59. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C - A research video collection. In: International Conference on Multimedia Modeling, pp. 349–360. Springer (2019). URL https://doi.org/10.1007/978-3-030-05710-7_29
60. Schall, K., Barthel, K.U., Hezel, N., Jung, K.: Deep aggregation of regional convolutional activations for content based image retrieval. In: International Workshop on Multimedia Signal Processing, pp. 1–6. IEEE (2019). URL https://doi.org/10.1109/MMSP.2019.8901787
61. Schoeffmann, K.: Vbs 2021 overview. URL https://www.youtube.com/watch?v=8Kg_5BQon9I&t=587s
62. Schoeffmann, K.: Video browser showdown 2012-2019: A review. In: Conference on Content-Based Multimedia Indexing, pp. 1–4. IEEE (2019). URL https://doi.org/10.1109/CBMI.2019.8877397
63. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: ASTER: An attentional scene text recognizer with flexible rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(9), 2035–2048 (2019). URL https://doi.org/10.1109/TPAMI.2018.2848939

64. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015). URL http://arxiv.org/abs/1409.1556
65. Smith, R.: An overview of the tesseract OCR engine. In: International Conference on Document Analysis and Recognition, pp. 629–633. IEEE (2007). URL https://doi.org/10.1109/ICDAR.2007.4376991
66. Soucek, T., Lokoc, J.: Transnet V2: an effective deep network architecture for fast shot transition detection. CoRR **abs/2008.04838** (2020). URL https://arxiv.org/abs/2008.04838
67. 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, pp. 441–447. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_42
68. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Conference on Computer Vision and Pattern Recognition, pp. 2818–2826. IEEE (2016). URL https://doi.org/10.1109/CVPR.2016.308
69. Tan, W.R., Chan, C.S., Aguirre, H.E., Tanaka, K.: Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. International Conference on Image Processing pp. 3703–3707 (2016). URL https://doi.org/10.1109/ICIP.2016.7533051
70. 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, pp. 490–495. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_50
71. Van De Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Transactions on Image Processing **18**(7), 1512–1523 (2009). URL https://doi.org/10.1109/TIP.2009.2019809
72. Veselý, P., Mejzlík, F., Lokoč, J.: Somhunter V2 at video browser showdown 2021. In: International Conference on Multimedia Modeling, pp. 461–466. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_45
73. Wu, J., Ngo, C.W.: Interpretable embedding for ad-hoc video search. In: International Conference on Multimedia, pp. 3357—-3366. ACM (2020). URL https://doi.org/10.1145/3394171.3413916
74. Wu, J., Nguyen, P.A., Ma, Z., Ngo, C.W.: Sql-like interpretable interactive video search. In: International Conference on Multimedia Modeling, pp. 391–397. Springer (2021). URL https://doi.org/10.1007/978-3-030-67835-7_34
75. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Conference on Computer Vision and Pattern Recognition, pp. 5987–5995. IEEE (2017). URL https://doi.org/10.1109/CVPR.2017.634
76. Ye, G., Li, Y., Xu, H., Liu, D., Chang, S.F.: Eventnet: A large scale structured concept library for complex event detection in video. In: International Conference on Multimedia, pp. 471—-480. ACM (2015). URL https://doi.org/10.1145/2733373.2806221
77. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: Varifocalnet: An iou-aware dense object detector. In: Conference on Computer Vision and Pattern Recognition, pp. 8514–8523. IEEE (2021)
78. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1452–1464 (2018). URL https://doi.org/10.1109/TPAMI.2017.2723009