On Influential Trends in Interactive Video Retrieval: Video Browser Showdown 2015-2017

Jakub Lokoč, Werner Bailer, Klaus Schoeffmann, Bernd Muenzer, George Awad

***Abstract*—The last decade has seen innovations that make video recording, manipulation, storage and sharing easier than ever before, thus impacting many areas of life. New video retrieval scenarios emerged as well, which challenge the state-of-the-art video retrieval approaches. Despite recent advances in content analysis, video retrieval can still benefit from involving the human user in the loop. We present our experience with a class of interactive video retrieval scenarios and our methodology to stimulate the evolution of new interactive video retrieval approaches. More specifically, the Video Browser Showdown evaluation campaign is thoroughly analyzed, focusing on the years 2015-2017. Evaluation scenarios, objectives and metrics are presented, complemented by the results of the annual evaluations. The results reveal promising interactive video retrieval techniques adopted by the most successful tools and confirm assumptions about the different complexity of various types of interactive retrieval scenarios. A comparison of the interactive retrieval tools with automatic approaches (including fully automatic and manual query formulation) participating in the TRECVID 2016 Ad-hoc Video Search (AVS) task is discussed. Finally, based on the results of data analysis, a substantial revision of the evaluation methodology for the following years of the Video Browser Showdown is provided.**

***Index Terms*—Interactive video retrieval, video browsing, content-based methods, evaluation metrics.**

I. Introduction

With the growing amount of video data in both professional and consumer applications, the past decade has seen a growing need for effective approaches to video search and retrieval, even on large-scale video datasets. Benchmarking initiatives such as TRECVID [1] and MediaEval [2] have been instrumental in establishing evaluation methods and fostering research in automated video retrieval [3], [4], [5], [6], [7], [8]. Despite these advances, many video retrieval problems are still challenging and can benefit from a human user in the loop. Work on interactive video search builds on the automatic methods, and focuses on topics such as making effective use of the selections of the human user and on intuitive and efficient user interfaces. For example, different dimensions of similarities are used in [9] to form threads, along which the user can explore the video database. In [10] similarity is exploited to find video segments that are related to those already identified as relevant by the user. [11] use external information such as Wikipedia to infer different facets of the query. Segments of the video are then aligned to these facets, with the resulting structure serving as a hierarchy for browsing. Other work shows that the presentation style of keyframes (e.g., 3D vs. 2D) can make a difference for interactive search [12]. Evaluation of such interactive search tools is more challenging than for automatic tools since it must also involve users.

The Video Browser Showdown (VBS) is an annual live video search competition, organized as a special session at the *International Conference on MultiMedia Modeling* (MMM) since 2012 [13]. During the session, international researchers evaluate and demonstrate the efficiency of their interactive video retrieval tools on a shared dataset (for 2017 the TRECVID IACC.3 data set with 600 hours of video content was used). In the live sessions, the researchers compete in *known-item search* (KIS) and *ad-hoc video search* (AVS) tasks (see Section III) in front of an audience. The aim of the Video Browser Showdown is to evaluate video browsing tools in a highly competitive setting, i.e., in direct comparison to search systems of other participants. It aims to foster research on interactive video retrieval tools [14], [15], [16], [17] that allow for content-based search in large video collections, a topic with increasing importance due to the ubiquity of video data in the digital universe. Instead of focusing only on the automated retrieval performance (e.g., "*query by text and browse the first few results"*), it addresses highly interactive search systems [18], which focus on the human in the loop [19] and which are able to reduce shortcomings of automatic visual content retrieval due to many flexible search features [20]. The VBS also motivates approaches with high usability (by an explicit *Novice Run*) to reduce the usability gap [21]. Last but not least, the VBS is an engaging and entertaining event at the MMM conference, which brings together researchers working in the field of multimedia retrieval to discuss and compare their latest approaches. For the conference audience, it showcases state-of-the-art video retrieval tools and allows them to look behind the scenes and even try different approaches themselves.

In this paper, we discuss the VBS of the last three years (2015-2017), evaluate the achieved performance (similar to the review presented in [22]) and summarize the best working approaches. More importantly, we revisit the rationale of the VBS, discuss its rules and how we can achieve fair scoring. Based on our findings, we propose changes in the organization of the VBS event in the next few years and highlight promising content-based search features for future video retrieval tools to incorporate to successfully participate in this challenging and entertaining competition.

J. Lokoč is with Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

W. Bailer is with JOANNEUM RESEARCH, DIGITAL, Graz, Austria

K. Schoeffmann and B. Muenzer is with Alpen-Adria-Universität Klagenfurt, Austria

G. Awad is with Information Access division, Information Technology Laboratory, National Institute of Standards and Technology (NIST), USA

The paper is organized as follows. In Section II, we discuss the idea of interactive video retrieval and its advantages over automatic *search-and-browse-results* approaches. The video browser showdown evaluation campaign and its objectives for 2015-2017 are presented in Section III. The selected top-ranked interactive video retrieval tools used in the analysis are presented in Section IV. Section V presents evaluation results and Section VI discusses these results and proposes some changes for future iterations of the VBS evaluation campaign. Section VII discusses trends and future challenges. Finally, Section VIII concludes the paper.

II. Interactive video retrieval

A. In defense of interactive video retrieval

The ability of a video retrieval system to handle information needs of its users depends on several factors. The data representation (data model) used in the system determines whether users can employ keyword search and/or query-by-example/sketch approaches. Given a data model, the effectiveness and efficiency of the utilized retrieval model determine the set of addressable user queries. A traditional approach to handle the information needs is automatic retrieval [23], [24], where users express all their intents in a (potentially complex/multimodal) query and the system returns top n most relevant scenes to the user. The systems are trained with respect to a benchmark dataset, often optimizing a precision-based measure. Considering a set of involved retrieval models $\mathcal{M}$ with parameters Θ producing a ranked list $L_{\mathcal{M},\Theta,q_i}$ for each query q_i from the set of benchmark training queries Q (i.e., with a defined relevant answer), the systems usually try to optimize (with a regularization preventing over-fitting) the Mean Average Precision

$$\arg\max_{\Theta} \sum_{i=1}^{|Q|} \frac{\sum_{k=1}^{n} \mathrm{rel}(L^k_{\mathcal{M},\Theta,q_i}) \mathrm{Pr}(L^k_{\mathcal{M},\Theta,q_i})}{nrel_{q_i} \cdot |Q|}, \qquad (1)$$

where n represents the number of returned frames/scenes, $nrel_{q_i}$ the number of relevant items in the whole collection for query q_i, and the inner sum evaluates precision at rank k, if the item at rank k is relevant ($\mathrm{rel}(\cdot) = 1$). Note that in cases where relevance judgments are incomplete, Inferred Average Precision (infAP) can be used to obtain a robust approximation [25].

However, in the video retrieval domain the automatic retrieval systems are still far from effective for many search tasks and queries. As shown in Table I, the top performing teams of the TRECVID Ad-hoc Video Search (AVS) task using manually assisted and automatic systems in the last two years reached just about 21% maximum Inferred Average Precision.

A summary of general approaches by teams at TRECVID [1] shows that most teams relied on intensive visual concept indexing, leveraging on past semantic indexing tasks and used popular datasets for training such as ImageNet [26]. Deep learning approaches dominated teams' methods, using pre-trained models [27], expansion and/or transformation approaches to map concepts to queries. Concept score fusion was investigated by most teams to combine useful results that satisfy the queries [28]. Other approaches investigated video to text and unified text-image vector spaces [29].

System Type	Year	Team	InfAP
Manually Assisted	2016	Waseda	0.177
Manually Assisted	2016	kobe_nict_siegen	0.047
Manually Assisted	2016	IMOTION	0.047
Manually Assisted	2016	vitrivr	0.044
Manually Assisted	2016	VIREO	0.044
Fully automatic	2016	NII_Hitachi_UIT	0.054
Fully automatic	2016	ITI_CERTH (VERGE)	0.051
Fully automatic	2016	INF	0.040
Fully automatic	2016	VIREO	0.038
Fully automatic	2016	MediaMill	0.034
Manually Assisted	2017	Waseda_Meisei	0.216
Manually Assisted	2017	VIREO	0.164
Manually Assisted	2017	FIU_UM	0.147
Manually Assisted	2017	ITEC_UNIKLU	0.102
Manually Assisted	2017	kobe_nict_siegen	0.089
Fully automatic	2017	MediaMill	0.206
Fully automatic	2017	Waseda_Meisei	0.159
Fully automatic	2017	VIREO	0.120
Fully automatic	2017	ITI_CERTH (VERGE)	0.095
Fully automatic	2017	EURECOM	0.094

TABLE I
Top 5 scoring teams at TRECVID Ad-hoc Video Search task (AVS) 2016-2017. Results are based on the ground truth from pooled TRECVID submissions.

Despite breakthroughs in end-to-end deep learning in image classification [30], [31], [32] and captioning [33], sufficiently large and error-free training data are still not available for effective learning of models for complex and variable video data (except in specific sub-domains). It becomes a classical chicken-and-egg problem: in the absence of an army of annotators, sufficient training data can only be collected by the system that needs the training data. Interactive video retrieval [18] represents a promising solution to break this cycle because it benefits from human-machine cooperation. While machines have proved to be very efficient for particular tasks, human intuition provides a high-level control over particular retrieval decisions and actions. This leads to a user-centric architecture [19], where query processing is split into several steps supported by interactive visualizations [34], [35], [36], [37]. In each step, the human intuition selects what action is applied next to approach the search scene. The interaction also provides relevance feedback, which can be used to temporarily narrow a potential semantic gap or effectively search for the mental picture [38], sometimes extremely quickly [39], [40], [41]. Incorporating user feedback and interactive visualizations has also been recommended in other areas, for example visual data mining [42], knowledge generation [43] or human-centered machine learning [44].

Another argument for interactive video retrieval is its ability to express widely varying search intents with respect to video content. When either the user is not able to sufficiently describe the search target (e.g., details of the face of a previously seen person) or the provided description is so abstract that it matches a large portion of the database (e.g., find a face in news videos), human interaction helps the search system to find relevant results quickly.

B. Taxonomy of interactive video retrieval tasks

Various interactive video retrieval task types are emerging both at competitions and in related literature. All the tasks expect that users are able to recognize a displayed correct result (i.e., to link the result to the mental query residing in his/her mind). In order to better specify/restrict interactive video retrieval tasks addressed by a competition, we present a general taxonomy determined by two properties – the cardinality of the result and the specificity of mental query initialization. Based on the cardinality, we distinguish two cases of search tasks (labels target/class taken from [38])

- *Target* – users search for exactly one particular scene. For example, a moment in a surgery video, a particular shot for video compilation, an event in the personal video collection or just one video that will entertain the user.
- *Class* – users search for multiple instances of a more abstractly defined scene. For example, all suspicious situations from street cameras, or all dancing persons in the news videos.

From the most specific to most abstract forms, four cases of mental query initialization can be considered

- *Example* – users have an ideal instance of the mental query available. For example, a distorted photo/shot of a searched scene or a picture of a target person.
- *Visual* – users search for a previously observed scene just based on their memory. For example, after inspection of medical surgery videos, film production and video editing (e.g., compilation), life logging or videos taken by car hood cameras.
- *Textual* – only a detailed description of a searched scene (e.g., by the witness of an accident) is available and video retrieval experts try to find the scene corresponding to the description.
- *None* – in the beginning, users just explore a video collection (e.g., for fun) by browsing and gradually identify the scenes of interest after some time. Note that this interesting use case is very difficult to simulate and evaluate. To the best of our knowledge, it is not considered in benchmarks/competitions.

Based on these two properties, eight categories of interactive video retrieval tasks can be organized in a two dimensional taxonomy table defined as $\{Target, Class\} \times \{Example, Visual, Textual, None\}$.

C. Interactive video retrieval evaluation

The interactive video retrieval option, often highly user centered, also causes additional difficulties in the evaluation of effectiveness. An automatic evaluation of complex interactive video retrieval systems is a difficult task to manage due to necessary simulations of (hard to predict) behavior of real users in corresponding complex interaction scenarios. Nevertheless, suitable user simulation approaches for specific subparts of the systems can be managed to identify potential bottlenecks, promising initial settings, or retrieval strategies [45]. Once an interactive video retrieval system is fine-tuned by a suitable set of simulations, evaluations involving real users are performed to test the system and identify potential usability bottlenecks. Options to the evaluation of interactive retrieval systems include surveys, analysis of system logs, question answering and indirect evaluation, i.e., measuring the effectiveness in performing specific tasks [46]. The latter approach has the advantage that at least part of the judged data set can be prepared and reused. Nonetheless, evaluations are still time-consuming and the number of participating users is usually low. Hence, in order to further compare many different tools given a unified set of tasks for interactive video retrieval, evaluation campaigns are organized.

D. History of interactive video retrieval evaluation campaigns

Following the realization that benchmarking IR tasks needed to scale up in size in order to be realistic, the Text Retrieval Conference (TREC) initiative began in 1991 [47]. It set out initially to benchmark the ad hoc search and retrieval operation on text documents and over the years spawned over a dozen or more IR-related tasks including video related tasks. One of the evaluation campaigns which started as a track within TREC in 2001 but spawned off as an independent activity after two years is the video data track, known as TRECVID. The motivation was an interest at NIST in expanding the notion of information in IR beyond text and the observation that it was difficult to compare research results in video retrieval because there was no common basis (data, tasks, measures) for scientific comparison. In its first set of evaluated tasks, TRECVID in 2001 benchmarked systems on ad-hoc search, instance search [48] and shot boundary detection where manual/interactive runs were an option for systems. Since then, other video retrieval tasks have been introduced and evolved such as high-level feature extraction (re-branded to semantic indexing later) [49], video summarization, video copy detection, multimedia and surveillance event detection, concept localization, video hyperlinking and most recently video to text description [1].

Since the TRECVID community developed very complex and powerful systems for solving the tasks mentioned above, they were perfectly suited to showcase the state-of-the-art in video retrieval – and the VideOlympics [50] were born. The VideOlympics competition was a video retrieval showcase performed at the ACM International Conference on Image and Video Retrieval (CIVR [51]) from 2007 to 2009 and it was the inspiration for the Video Browser Showdown. In the VideOlympics, around nine participating TRECVID teams had to solve ad-hoc video search tasks on a 160 hours video collection (around 80,000 shots). An evaluation server presented a scoreboard showing correct and wrong submissions for each team, according to the TRECVID ground truth. In an Olympic spirit, participation was more important than winning and several winners were awarded with a Golden Retriever award, partially by voting of the audience (most impressive interface, public's favorite, etc.). When the VideOlympics were discontinued after 2009, the Video Browser Showdown adopted the idea of this live video search competition. However, in contrast to the VideOlympics, interactive browsing was more important than content-based retrieval in the first two showdowns [22],

[52]. In the third year, the VBS started to evaluate video search in a collection (which was increased in terms of content size over the years) as well as issuing textual KIS queries.

III. Video browser showdown 2015-2017

In this section, three years of the new unconstrained retrieval era of VBS are summarized, focusing on evaluation scenarios, objectives and metrics.

A. Evaluation scenarios

According to the taxonomy presented in Section II-B, the VBS competition focuses on three different task categories. Each category is represented by one well-established task type.

a) Visual known-item search (KIS) task: – users search for a randomly selected short scene (20s), played in a loop using a data projector for a given time limit. The loop ensures that the users know what they search for. The scenes are selected to be visually unique in the dataset. However, the scenes may be composed of multiple shots, and there may be similar individual shots elsewhere in the dataset, but never in the same temporal composition. This task represents the *Target Visual* category.

b) Textual known-item search (KIS) task: – users search for a randomly selected short scene (20s), described by a text presented using a data projector. Like for the visual known-item search task, the target scenes are selected to be unique in the data set. In addition, scenes consisting of a single shot or few related shots are selected, in order to enable a compact and yet unambiguous description. All key elements of the scene that discriminate it from similar scenes in the dataset are included in the description. This task represents the *Target Textual* category.

c) Ad-hoc search (AVS) task: – users search for multiple video segments that match a short textual description presented via data projector. This textual query may refer to persons, objects, activities, locations, etc. and combinations of the former. This task, introduced in VBS 2017, represents the *Class Textual* category.

The competing teams have to find and submit a keyframe from the searched scene. Assuming that each frame is identified by its number corresponding to its time position in a video, the correct result of the KIS search task can be represented as a triplet

$$KIS\_task_i = \langle video^i, frame^i_{start}, frame^i_{end} \rangle,$$

where $video^i$ is a unique video identifier and $frame^i_{start} < frame^i_{end}$ represent frame number based boundaries of the searched scene. A submission of team$_j$ in task$_i$ is defined as a triplet (vid^i, f^i_j, t^i_j), where f^i_j represents the submitted frame number from a video vid^i and t^i_j represents the corresponding elapsed search time. The submission of team$_j$ is defined to be correct for KIS_task_i if $video^i = vid^i$ and $f^i_j \in [frame^i_{start}, frame^i_{end}]$. The submission is defined to be incorrect/wrong otherwise.

The correct result of an AVS search task is defined as a set of triplets, each corresponding to one correct scene

$$AVS\_Task = \{\langle video^i, frame^i_{start}, frame^i_{end} \rangle\}_{i=1}^{k}$$

In order to completely solve the *AVS* search task, a team has to provide a set of k submissions, with one correct submission for each triplet in AVS_Task.

B. Evaluation objectives

Since 2015, the main objective of the VBS competition has been to identify interactive video retrieval tools capable of efficiently solving both visual and textual known-item search tasks without any restrictions on the developed tools. Moreover, with the unlimited arsenal of possible multimedia retrieval approaches, only search in a large video archive (increasing every year) was considered as a real challenge for the competing tools. The objective of no restrictions went so far in VBS 2015 that even cameras were allowed for visual KIS tasks (addressing also the *Target Example* category). However, taking pictures/shots is not always possible and it also leads to the preference towards automatic retrieval approaches already covered by computer vision and machine learning benchmarks. Therefore, at VBS 2016 and 2017 cameras were prohibited once again.

Whereas interactive known-item search in short videos challenges mainly retrieval efficiency, for archive search an obvious new objective is effectiveness, determined by the number of solved tasks within an acceptable time. But this causes a new evaluation challenge. When there is no clear winner, should the VBS competition appreciate efficiency of less effective tools or effectiveness of less efficient tools? Although the effectiveness usually matters, the exhausting and also time-limited nature of VBS evaluation affects the preference direction. In order to evaluate sufficiently enough KIS tasks within a limited time frame at the MMM conference, all unfinished search sessions have to be terminated after a relatively short time limit $t_L \in R^+$ (five minute limits were mostly used). The requirement of such a short time limit leads to the preference of highly efficient tools based on querying/filtering, even though other tools relying on more systematic exploration could be more effective for a higher time limit. Note that the successful tools at VBS often incorporated various query initialization approaches in order to quickly localize candidates for inspection (see Section IV).

Another issue related to the limited time frame for VBS is an evaluation with novice users selected from the audience. The usability gap represents a traditional evaluation objective of VBS. However, according to our experience the skills of selected novice users can vary a lot. While novice sessions were considered in the team ranking at VBS 2015, at VBS 2016 novice sessions were evaluated but not considered in the ranking (which affected the ordering of the teams). At VBS 2017, the time frame for the competition was strongly restricted and so the novice sessions were not evaluated. Both novice session and preference dilemma issues are addressed in Section VI.

C. Evaluation metrics for KIS tasks

Generally, a well performing interactive video retrieval tool should be competitive in basic requirements, reporting high recall, low number of false hits and low search times. So far, the VBS competition has considered a single formula scheme for all KIS tasks assigning a score $s_j^i \in N$ to each team$_j$ for each task$_i$. The formula scheme is defined for each KIS task$_i$ as a binary function

$$f_{KIS}^i : R_0^+ \times N \to N,$$

assigning the score to each team$_j$ for its correct submission time $t_j^i \in R_0^+$ and the number of preceding wrong submissions $ws_j^i \in N$. The score of team$_j$ in known-item search is defined as the sum of its scores over all k performed KIS tasks

$$score_{team_j}^{KIS} = \sum_{i=1}^{k} f_{KIS}^i(t_j^i, ws_j^i)$$

The score function f_{KIS}^i is defined to be zero if the correct submission time t_j^i of team$_j$ is higher than the task$_i$ time limit $t_L^i \in R_0^+$ (not available t_j^i is set to $+\infty$). For all received correct submission times $t_j^i < t_L^i$, three components are employed – a constant score $s_C \in [0, 100] \subset N$, a decreasing time-score function $f_{TS} : R_0^+ \to R$ and a wrong submission penalty function $f_{WSP} : N \to N$. The final score is defined as

$$f_{KIS}^i(t, ws) = \left\lceil \frac{(s_C + (100 - s_C) \cdot f_{TS}^i(t))}{f_{WSP}(ws)} \right\rceil$$

where the numerator assigns the score for solving the task (see an example function in Figure 1), while the denominator penalizes for wrong submissions. The purpose of s_C is to provide a time-independent score reward for solving the task before the time limit t_L^i, while the decreasing function $f_{TS}^i(t)$ function differentiates between fast and slow successful tools. Due to potential KIS task ambiguity issues, teams are allowed to submit more submissions with an immediate feedback whether the submission is correct. The wrong submission penalty function f_{WSP} penalizes for the misuse of this opportunity.

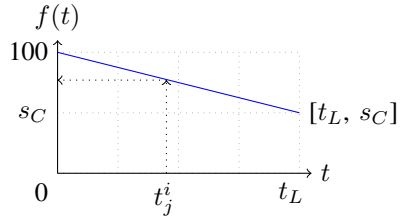

Fig. 1. The function used to assign a score for a correct KIS task submission time t_j^i in interval $[0, t_L]$, before a wrong submission penalty is applied.

At VBS 2015 and 2016, the score function was defined as

$$f_{KIS}^i(t, ws) = \left\lceil \frac{(50 + 50 \cdot (t_L^i - t)/t_L^i)}{\max(1, |ws| - 1)} \right\rceil$$

where s_C, set to 50, $f_{TS}^i(t)$, was a linearly decreasing function $(t_L^i - t)/t_L^i$ employing the task time limit constant t_L^i and the wrong submission penalty function f_{WSP} allowed up to two wrong submissions for free. At VBS 2017, another time score function $f_{TS}^i(t) = (t_L^i - t)/(t_L^i - t_M^i)$ was tested, where $t_M^i = \min(t_1^i, \cdots, t_{|teams|}^i)$. The new function always assigns 100 points to the team with the first correct submission (before wrong submission penalty is applied) and thus provides a higher reward with respect to unsuccessful teams in task$_i$. The new function also causes a steeper slope for more difficult tasks, which introduces higher differences for successful teams. However, the new $f_{TS}^i(t)$ had no effect on the resulting ordering in 2017 and it will not be considered in the future evaluations. A more detailed analysis and discussion on the score formulas are presented in Section VI.

D. Evaluation metric for AVS tasks – pilot setting

The evaluation metric for AVS tasks for team$_j$ and task$_i$ considered only the sets of correct C_j^i and incorrect I_j^i submissions. As the employed TRECVID-based ground truth is created from pooled submissions and thus cannot be expected to be complete, a live judge was necessary to assess submissions not present in the ground truth. The time limit t_L affected only the time for the entire search session. The scores from all the AVS sessions were added and the overall score from the AVS tasks was added to the overall score from KIS tasks.

$$score_{team_j}^{AVS} = \sum_{i=1}^{k} f_{AVS}^i(C_j^i, I_j^i),$$

For one task, the score function was defined as:

$$f_{AVS}^i(C, I) = \left\lceil \frac{(100/C_{max}) * |C|}{\max(1, |I| - 1)} \right\rceil,$$

where C_{max} is the maximum number of correct submissions over all teams for the given task. This means that the team finding the most correct shots determines the number of points per correct submission. The penalty function was naively adopted from the KIS scoring function. According to the results presented in Section V, the pilot AVS score formula did not affect the overall ranking (except the last two teams) and did not distribute score points well in several situations. One of the major problems is the penalty function, which turned out too be far too onerous. We discuss a new AVS score formula in Section VI.

IV. Scoring tools at VBS 2015-2017

The results of the VBS competition highlight promising trends in interactive video retrieval. Usually, the top ranked tools influence other teams that adopt the successful approaches in new versions of their tools. As a result, more and more complex tools participate in the competition, where often a few most up-to-date tools with similar performance compete for the best score. Although many interesting and novel approaches have been presented during the last three years, in this section only the top three performing tools are presented in more detail for each year and their influential ideas are summarized. Since many of the tools evolve in time (including new tool names), Table II presents the names of the

teams used during the competition. In the analysis section, we will use the team names to refer to the corresponding developed systems.

Year	2015	2016	2017
DB	100h	250h	600h
1st	SIRET [53]	HTW [54]	IMOTION [55]
2nd	IMOTION [56]	ITEC-UU [57]	ITEC-UU [20]
3rd	UU [40]	SIRET [58]	SIRET [59]
4th	HTW [60]	VERGE [61]	VIREO [62]
5th	NII-UIT [63]	UoS [64]	NII-UIT [65]
6th	VERGE [66]	IMOTION [67]	VERGE [68]
7th		JRS [69]	
not ranked	iAutoMotion [70]		

TABLE II
SCORING TEAMS AT VBS 2015-2017.

In the following, we present key features of the top three performing tools from each year. Although the remaining teams presented interesting approaches briefly presented in Section IV-F, due to lack of space we refer the reader to the corresponding publications for more detailed descriptions.

A. SIRET team

Based on the success of the SBVB tool introduced by SIRET at VBS 2014 [71], an enhanced version [53] was presented at VBS 2015. The new version again relied mostly on the interactive color-sketch interface with an improved retrieval model and compacted visualization of the results. The tool also supported query by camera-photo, which was allowed only at VBS 2015. At VBS 2016, the tool received two main improvements [58] – query by dominant edges was integrated into the color-sketching canvas (inspired by the IMOTION team [56]) and similarity search using DCNN features was added. Although the tool was competitive in visual KIS tasks (mainly the color-based retrieval and similarity search), the tool was not able to solve any textual KIS tasks due to improper use of query initialization methods. Hence, the purely content-based retrieval interface was extended by two query initialization options in 2017 [59] – query-by-external-example (e.g., from Google images) and query-by-keywords. For the automatic annotation of video keyframes, the pretrained VGG network using 1000 concepts and hypernym expansions from WordNet were used. The tool also adopted a display-organization approach inspired by the HTW team [60]. For a dynamically evaluated kNN query, the tool sorts the results in a grid layout on demand. Surprisingly, in 2017 the SIRET team still solved many tasks using the interactive color-sketch model from year 2015 that proved to be orthogonal/competitive to the state-of-the-art interactive video retrieval tools (often based on DCNN approaches) for the VBS 2017 settings.

B. IMOTION team

At VBS 2015, the IMOTION team presented a versatile sketch-based tool [56], supporting retrieval using various classes of low-level features – shot position, global color features, regional color/edge features, and motion features. Two types of additional high-level features were obtained from deep convolutional neural networks, considering spatial and/or temporal information. The system supported three modes of queries – query-by-sketch, query-by-example, and motion queries. A relevance feedback approach was incorporated to refine the query results for a given search session. In 2016, the tool [67] was not in the cluster of top performing tools (6th place), probably because the camera-photo option was disabled. Note that the team presented also a separate system based on recorded video [70], which was considered just as a baseline for visual KIS tasks (out of the competition ranking). In 2017, the improved version of the tool [55] achieved the best score and won the competition. In this latest iteration, the system added more detectors and a new interface for semantic query specification, improving the filtering power of the tool. Deep neural networks were used to obtain more semantic concepts, for similarity search and for display organization. The browsing capabilities were extended by a graph-based result navigation interface (inspired by the HTW team). A new underlying storage system ADAM$_{pro}$ [72] was utilized for large scale retrieval.

C. UU team

Unlike the other teams at VBS 2015, the UU team presented a simple and very intuitive video browsing tool for mobile devices, completely excluding video analysis and query processing [40]. The tool employed an optimized small screen storyboard layout (625 keyframes on one screen, one extracted keyframe per second) utilizing touch screen gestures to browse. Although exhaustive human computing was required to browse the whole collection (100 hours of video = 576 screens), the tool outperformed all the tools in the expert run. Especially the textual tasks were effectively solved by the expert user who gradually learned the contents of the collection. However, in the novice run the tool was outperformed by most of the other tools. Nevertheless, the approach has proven to be highly competitive for one hundred hour video collection in the comparison with complex desktop PC systems. In the following years, this approach became a part of the collaborative tool developed jointly by ITEC and UU teams [57], [20] (see Section IV-E).

D. HTW team

In 2015, the HTW team placed fourth place with a score close to the top three tools. The proposed tool [60] used a graph-based browsing system extending a previously proposed ImageMap system for hierarchical visual browsing of large image collections. The graph was created using an image and edge swapping heuristic in the 2D grid. The tool employed low-level visual features and keywords propagated from annotated known images to unknown frames. The tool supported browsing and search modes. The browsing mode used a fractal tree map and a sorted graph projection enabling graph exploration. The search mode supported query-by-example, color, or sketch to localize promising frames for inspection. In 2016, the team extended their approach by adding semantic features learned from deep convolutional neural networks to improve the graph quality and a new projection method to

preserve complex relationships in the 2D grid [54]. Using its own sketch feature vectors, the new version also adopted the timed-sketch approach successfully applied by the SIRET team [71], [53]. With the new version of the tool, the team has won VBS in 2016 but did not participate at VBS 2017.

E. ITEC-UU team

The ITEC-UU team [57], [41] combined the idea of the mobile storyboard, presented by the UU 2015 team, with a desktop-based video retrieval tool. In their collaborative approach two users work side-by-side with different interfaces: one with a storyboard layout on a mobile device, and the other with a desktop-based video retrieval tool. The mobile storyboard was almost identical to the one used by the UU team in 2015, but it was connected to the retrieval tool and exchanged information about already inspected segments (in the storyboard) and retrieved results of queries (in the desktop tool) to dynamically reorganize content. The desktop tool allowed retrieving shots according to visual concepts detected with a deep neural networks trained with 1000 classes from ImageNet [30]. Moreover, it also included a sketch-based search function, relying on feature signatures [73], [74], computed for keyframes of shots and compared with the signature matching distance [75]. The tool was completely redesigned for VBS 2017 [20], where an integrated web-based system was used, which contained several query features: a hierarchically organized map of keyframes that are arranged by similarity (inspired by the HTW system [60]), the storyboard, an HSV-based color filter, a concept-based search, and a query-by-web-example feature. Furthermore, this tool implemented several features for collaborative use with several searchers.

F. Sparkling ideas of the other tools

In addition to the detailed description of the most successful tools, we briefly present several promising/inspiring ideas implemented by the remaining tools. The UoS team [64] contributed a system relying on easy-to-use visual mental browsing. The system tries to incrementally learn the mental target using a Bayesian framework, while the users just select one out of eight displayed shots in each iteration. The NII-UIT team [63], [65] proposed a system that used multimodal search as well as the combination of multiple semantic concepts from different neural networks. The authors also proposed a sketch based search system to leverage automatically detected concepts and the spatial relations between them. The VIREO team [62] presented a human-in-the-loop enhancement of the system they used to participate in TRECVID. The authors incorporated concept screening, video re-ranking by highlighted concepts, relevance feedback and color sketch to refine result sets. The VERGE team [66], [61], [68] incorporated content-based analysis and retrieval modules such as video shot segmentation, concept detection, clustering, as well as visual similarity and object-based search. Similar to the other teams, the authors shifted their models to deep convolutional neural networks both for automatic annotation and similarity search. The tool also employed automatic query formulation and expansion using high level concepts. For visualization, clustering by color using Self Organizing Maps and embedding of CNN-based descriptors using t-SNE method were implemented. The JRS team [69] implemented an approach for iteratively narrowing down the content set by applying a sequence of filters, which allow selection of clusters based on a similarity range. The features used are motion activity, camera motion, global color features, VLAD signatures and object trajectories.

V. Analysis of results at VBS 2015-2017

A. KIS tasks at VBS

The results of all teams at all KIS tasks are visualized in Figures 2-4. Generally, the top scoring teams also have the highest number of correct submissions. Note that in 2015, the UU team (3rd) had one more correct submission with respect to SIRET (1st) and IMOTION (2nd) teams. However, the time to solve the task by the UU team was often more than $3/5$ of the time limit. Note also that the wrong submission penalty reduced a task score just three times in 2015 (no effect on the final order of teams), nine times in 2016 (penalties in four tasks for ITEC-UU pushed the team to the second place) and did not reduce the score at all in 2017 for KIS tasks.

One aspect we can observe over the last three years is that textual KIS tasks are much harder to solve than visual tasks. For example, in 2017 the top-3 teams could solve 76.2% of visual tasks, but only 28.6% of textual tasks. The numbers of 2016 and 2015 are similar but higher (except textual tasks in 2016), most probably due to the smaller dataset (2016: 83.3% vs. 23.3%, 2015: 81.0% vs. 50.0%). Also the search time for solved visual KIS tasks is much smaller: the top-3 teams required only 139.96s in average to solve them, but 195.21s for textual tasks (2016: 151.02s vs. 178.49s, 2015: 116.37s vs. 146.59s). These numbers also show that the search time of the top-3 teams continuously increased with the increasing size of the data set over the years for the textual KIS tasks. However, this is not fully true for the visual KIS tasks, where in 2017 the search time was lower, despite the larger data set. We hypothesize that this is a result of the ever-improving systems of teams participating for several years. Both aspects – the higher number of submissions for visual tasks, and the lower search times – are clearly visible in Figures 2-4, which visualize correct submissions for the last three years for the scoring teams (left part for experts (E) and right part for novices (N); lower is better in terms of search time). Please note that for 2016 there were no textual tasks for novices and 2017 no novices tasks at all.

B. AVS Task at VBS – first experience

In VBS 2017, the AVS 2016 queries and the corresponding dataset (IACC.3) were used in order to complement the task with a fully interactive version. Five random queries (out of 30) from the TRECVID 2016 AVS competition were selected for that purpose, while two additional AVS queries were specifically created for VBS 2017 (Table IV lists the queries). All but one team (namely SIRET) also participated in TRECVID 2016 AVS, so the first five AVS queries were not new to these teams. However, this provided no obvious

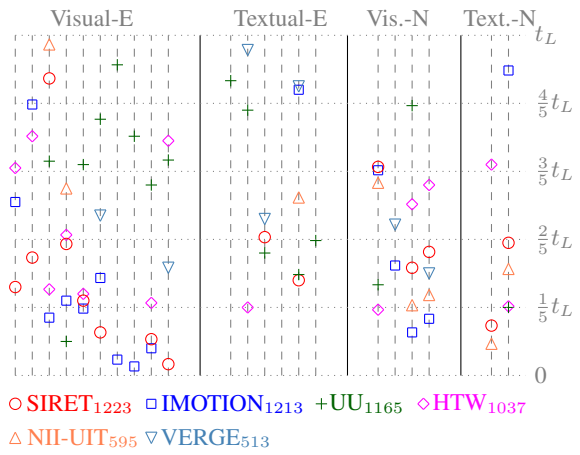

Fig. 2. Correct submission times of scoring teams at VBS 2015 in visual/textual KIS expert/novice sessions. The overall VBS score is next to the team label.

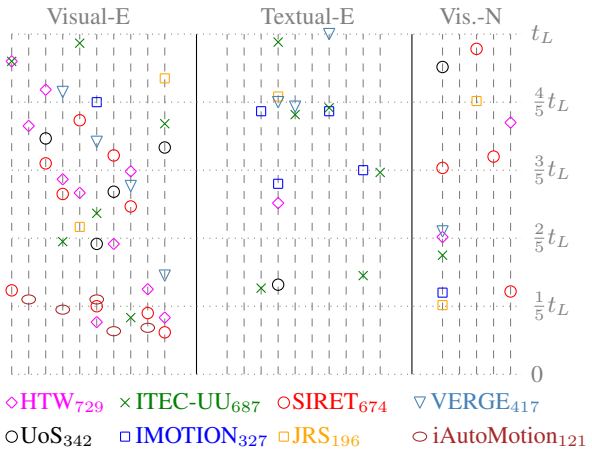

Fig. 3. Correct submission times of scoring teams at VBS 2016 in visual/textual KIS expert sessions and one visual KIS novice session. The overall VBS score is next to the team label (visual KIS novice session score is not counted).

benefit to them. SIRET and NII-UIT submitted significantly fewer results on average (16.71 and 15.85), while VERGE and ITEC submitted the most (38.71 and 47.00) – with a similar distribution of correct submissions, although the average precision of SIRET and NII-UIT (77.67% and 70.41%) was higher than for the other four teams (ranging from 58.95% to 66.52%). For the last AVS task ITEC submitted a whopping 166 submissions (average of 33.4 for the other teams), whereas only 97 of them were correct. These 166 submissions came from 42 different video files, and eight videos contained no correct submission. The ITEC team scored only 2 points in this task, due to the many wrong submissions (see Figure 5). For comparison, the SIRET team submitted a total of 14 results in this task, all of them correct, and received 15 points (see the top part of Table III). The reason for the high number of wrong submissions from the ITEC team was a different understanding of the query VBS_2 which led to the fact that all submissions showed close-up/faces, but not all of them contained talking faces. This aspect is clearly visible in Figure 5, which shows

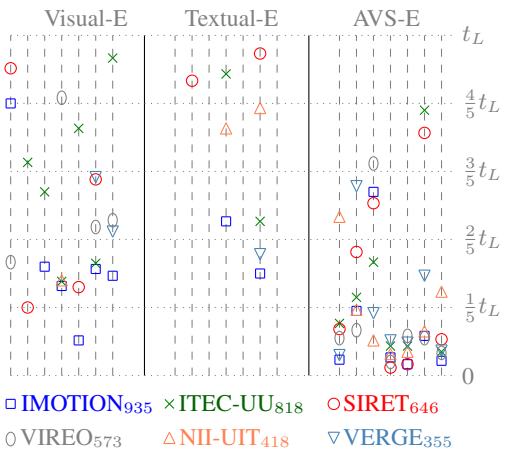

Fig. 4. Correct submission times of scoring teams at VBS 2017 for visual/textual KIS expert sessions (first two columns) and last column presents also the first correct submission times of AVS tasks. The overall VBS score is next to the team label.

that there were many wrong submissions for the last two AVS queries (VBS1 and VBS2) as well as for the TRECVID AVS query 505, which contained a rather ambiguous description. However, it is clear that the search performance of the ITEC team for task VBS2 was much better than the one from SIRET (97 correct submissions from 34 different videos vs. 14 correct submissions from 12 different videos). Therefore, we propose a new scoring function for AVS tasks in Section VI.

Generally speaking, the new AVS score function considers a form of precision at the recall level given by the set of correct submissions and the pool of all correct submissions of all teams. The function employs merging of shots in each video to quantize the recall axis. The results of two opposite extreme quantizations of shots (video or shot based recall) using the new AVS score function are presented in the middle and bottom part of Table III. In both cases, we observe that in VBS_2 task the ITEC team receives more points than the SIRET team. We may also observe that the SIRET team (focusing probably on high precision) is the best considering video-based recall and the worst considering the shot-based recall. As many of the correct submissions correspond to consecutive shots from one video, it provides easy recall gains. At the same time, teams should get points for finding unique shots from distinct parts of a video. Hence, we suggest to merge the consecutive shots and thus to obtain a compromise for the AVS scores (see Section VI).

A further lesson learned for future competitions is that it is crucial to have well ”calibrated” live judges. Their criteria for considering a submission as correct should already be defined and agreed on a priori. Otherwise, their interpretation of the query may be influenced by the chronology of submissions and may become inconsistent over time, which may adversely affect teams randomly. To prevent ambiguities, the teams could also ask few questions about the query before the session starts.

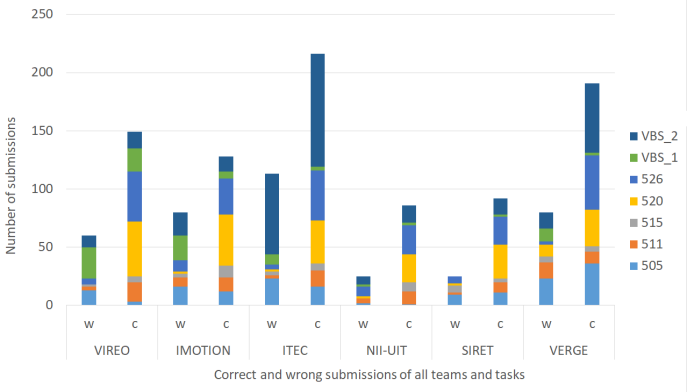

Fig. 5. Number of submissions in the AVS task at VBS 2017, aggregated by team and correctness.

	VIREO	IMOTION	ITEC	NII-UIT	SIRET	VERGE
VBS_2	2	1	2	3	15	5
VBS_1	4	2	2	10	10	1
526	23	8	31	8	11	50
520	100	94	79	52	62	8
515	50	50	30	80	6	13
511	50	11	42	22	53	5
505	1	3	3	3	4	5
SUM	230	169	189	178	161	87
VBS_2	5	7	31	13	15	28
VBS_1	20	12	20	22	33	4
526	23	24	20	18	20	28
520	42	49	8	40	48	57
515	13	18	17	37	8	7
511	23	14	17	16	28	4
505	3	25	11	2	32	38
SUM	129	149	124	148	184	166
VBS_2	5	4	35	6	7	26
VBS_1	40	7	4	4	7	2
526	22	15	22	12	12	25
520	59	54	45	29	35	33
515	11	24	13	22	4	9
511	26	15	21	16	14	10
505	1	10	13	1	11	38
SUM	164	129	153	90	90	143

TABLE III
THE TOP PART OF THE TABLE SHOWS THE ORIGINAL SCORES OF TEAMS AT AVS TASKS AT VBS 2017. THE MIDDLE PART SHOWS THE RESULTS OF THE NEW AVS SCORE FUNCTION CONSIDERING VIDEO BASED RECALL AND THE BOTTOM PART RESULTS OF THE AVS SCORE FUNCTION CONSIDERING SINGLE SHOT BASED RECALL. THE RESULTS ARE BASED ON THE EXTENDED GROUND TRUTH, INCLUDING THE ITEMS JUDGED DURING VBS.

C. Comparing AVS Task at VBS 2017 and TRECVID 2016

The TRECVID Ad-hoc Video Search task [1] models the use case of looking for video segments containing persons, objects, activities, locations, etc. and combinations of the former in an indexed video corpus, but without training on annotations for specific concepts. TRECVID runs two versions of an automatic task, either fully automatic (i.e., the system takes the textual query as input) or manually assisted (i.e., a user specifies a query to the system based on the query text). Of the 52 TRECVID 2016 AVS submissions, 22 were manually assisted submissions, while the other 30 were fully automatic.

Two important differences between TRECVID AVS and VBS submissions have to be considered. The first is that TRECVID submissions are ranked (i.e., the first shot is considered most relevant), while the VBS results are collected over time, but are not ranked by confidence in their correctness. The second is the number of returned shots. While for the TRECVID tasks, all participants return the maximum number of 1,000 shots, the average number of submitted shots in VBS is 29.64. For the five queries that were shared between TRECVID 2016 and VBS 2017, the average number of relevant shots is 291. The shared queries are listed in Table IV.

Topic ID	Text definition
505	Person holding a poster on the street at daytime
511	Destroyed buildings
515	Person jumping
520	Any type of fountains outdoors
526	A woman wearing glasses
VBS_1	Find shots out of an airplane windows, showing sky and clouds (and maybe part of a wing)
VBS_2	Find close-up shots of a talking face or mouth

TABLE IV
TRECVID 2016 AND VBS 2017 AD-HOC QUERIES. THOSE WITH NUMERIC IDS WERE COMMON BETWEEN TRECVID AND VBS.

For deciding about correctness of a submitted shot, VBS uses the ground truth provided by TRECVID. Unfortunately, this ground truth is not complete because it has been created based on (a subset of) the TRECVID AVS submissions. This means that VBS submissions may contain shots which have not been assessed yet, which necessitates a live judging mechanism. A nice side effect of this is the fact that the TRECVID ground truth is extended with new judgements. A further extension has been achieved in the course of an experiment at AAU (Alpen-Adria-Universität Klagenfurt), which used the same shared tasks. The extended ground truth set is publicly available[1]. In order to allow for comparison of the infAP, the scores from the TRECVID 2016 AVS tasks have been recalculated using the extended ground truth. Please note that the update of infAP is consistent with the TRECVID evaluation process.

For the TRECVID AVS submissions, we have infAP as a metric (calculated from the ranked list), while for the VBS submission we have precision and recall for the submitted set of items. For comparing the correlation of scores and ranks between tasks and teams, we use these original metrics. In order to enable a direct comparison, we use two approaches: simulating AP for the VBS submission, and calculating precision and recall for specific ranks for the TRECVID AVS submissions.

a) Comparing common topics between AVS and VBS: We compare the results for shared topics between AVS and VBS, i.e., whether the task difficulty in the fully automatic/manually supported and interactive scenarios is comparable. The raw data is provided in Tables V-VII. Note that we use the infAP and precision scores respectively, as those are the original results obtained. We consider the correlation of scores (and ranks) between the AVS and VBS results, thus the two types of metrics can be compared.

[1]http://www-nlpir.nist.gov/projects/tv2016/pastdata/extra.avs.qrels.tv16.xlsx

Task	Topic ID	Min	Median	Max
M	505	0.000	0.009	0.178
M	511	0.000	0.009	0.146
M	515	0.000	0.004	0.047
M	520	0.000	0.061	0.291
M	526	0.000	0.012	0.042
F	505	0.000	0.001	0.020
F	511	0.000	0.002	0.109
F	515	0.000	0.002	0.018
F	520	0.000	0.007	0.219
F	526	0.000	0.005	0.024
VBS	505	0.188	0.420	0.610
VBS	511	0.417	0.776	0.850
VBS	515	0.333	0.690	1.000
VBS	520	0.756	0.942	1.000
VBS	526	0.756	0.848	0.940

TABLE V
STATS OF COMMON TOPICS BETWEEN TRECVID AVS TASKS/TOPICS ACROSS ALL RUNS (INFAP, M : MANUALLY-ASSISTED, F : FULLY AUTOMATIC) AND VBS (PRECISION). THE RESULTS ARE BASED ON THE EXTENDED GROUND TRUTH, INCLUDING THE ITEMS JUDGED DURING VBS.

We determine the mean correlation of the scores obtained in each shared AVS and VBS topic. The mean correlation coefficient is -0.11, and the mean rank correlation coefficient is slightly positive at 0.15. The correlations between automatic and interactive tools are quite different for each topic. For example, for rank correlations of tasks 505 and 511 there is a clear positive correlation (> 0.7), while task 515 is uncorrelated and the other two are negatively correlated.

b) Comparing common teams' performance in AVS and VBS: We did a similar analysis for comparing the teams' scores. This mean correlation coefficient is 0.15 and the mean rank correlation coefficient is 0.20. While the VIREO and VERGE scores are positively correlated in terms of rank correlation, the ITEC_UU and IMOTION results are negatively correlated. One possible conclusion is that the VIREO and VERGE VBS tools reuse similar features, and thus perform more similar as in the AVS tasks, while the ITEC_UU and IMOTION tools for VBS seem to be more user-oriented, and thus are complementary to the AVS tools of these teams in some tasks.

c) Comparison using AP: The submission order in VBS does not express any ranking or confidence of the results. We thus make an estimate of the AP from all distinct permutations of the submitted result list. We calculate the AP of each of the permutations, and calculate the mean of these values, which we name "simulated AP". In fact, the calculation was not actually simulated, but the approach to analytically calculate this value is described in the appendix.

Figure 6 shows the comparison of infAP and simulated AP for the TRECVID fully automatic and manually assisted tasks and the VBS tasks. There is a correlation of success rate for the tasks across the two types of AVS and VBS runs, and task hardness is reflected in a similar way. For 505 and 511 the best manually assisted AVS results perform better, while for 526 the VBS results perform better. However, the median and bottom of the scores is significantly higher in the VBS submissions.

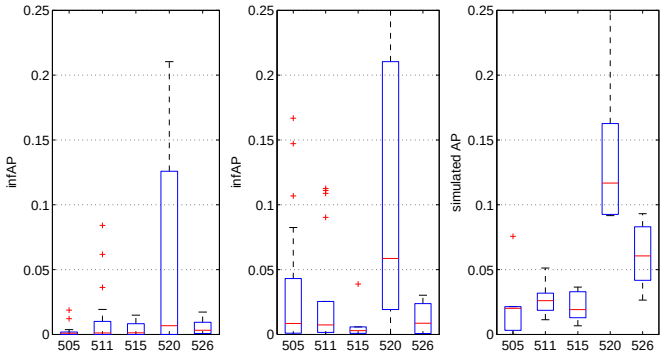

Fig. 6. Comparison of results AVS fully automatic (left) and manual (infAP) (middle), and simulated AP from VBS (right). The results are based on the extended ground truth, including the items judged during VBS.

d) Comparison using P and R: We compare precision and recall by calculating both for specific working points of the best TRECVID AVS submissions. More specifically, the precision at the recall level reached by VBS submissions is presented.

Figure 7 shows these results. While most TRECVID AVS submissions have lower precision at the same recall as the VBS submissions, there are some that outperform corresponding VBS submissions. Interestingly, this does not only include manually assisted, but also fully automatic ones. For one task, the best VBS submissions are outperformed, while for the other tasks, only some of the lower scoring VBS submissions are outperformed. We do not have detailed data from the tools to analyze the reasons. One hypothesis is that there are tasks, where the words in the query enable some automatic tools to generate a quite good results list, with many relevant items at the top of the list, while the users' queries in the interactive task captured only part of the relevant aspects and the tool operators focused on a riskier faster submission strategy without proper inspection of the shots (maybe due to time pressure). Furthermore, in several cases the subjective understanding/interpretation of the topic description by real users did not match the provided ground truth.

VI. WHERE ARE YOU HEADING VBS?

The video browser showdown competition will continue a successful cooperation with the MMM conference and TRECVID. The actual TRECVID dataset will be preserved to ensure a level of comparability with the state-of-the-art automatic video retrieval methods. However, based on our gained experience several aspects of VBS will be reconsidered.

A. Evaluation scenarios

All three evaluation scenarios will be preserved as they still challenge state-of-the-art interactive video retrieval tools. What will change are the scenario simulation settings that will be modified to reflect more realistic assumptions about the modeled situations.

For the visual KIS tasks, playing a randomly selected short scene in the loop simulates the assumption that the user undoubtedly knows what they search for. On the other hand,

Topic ID	Total TPs-O	Total TPs	Total FPs	Mean TP VIREO	Mean TP VERGE*	Mean TP ITEC_UU	Mean TP IMOTION	Mean FP VIREO	Mean FP VERGE	Mean FP ITEC_UU	Mean FP IMOTION
505	283	302	7129	15	26	8	38	985	974	584	962
511	223	289	5821	20	10	17	33	980	990	884	967
515	181	219	6654	16	13	6	14	984	986	700	986
520	171	178	6653	60	91	3	12	940	908	839	987
526	337	468	5907	78	52	4	70	922	948	667	930

TABLE VI
RELEVANT SHOTS STATS OF COMMON TEAMS AND TOPICS AT TRECVID AVS (*THE TRECVID TEAM NAME IS ITI_CERTH). TPS-O REPRESENT ORIGINAL TRECVID VALUES BASED ON THE GROUND TRUTH FROM POOLED TRECVID SUBMISSIONS. THE REMAINING COLUMNS ARE BASED ON THE EXTENDED GROUND TRUTH INCLUDING THE ADDITIONAL JUDGMENTS AFTER VBS SUBMISSIONS.

Topic ID	Mean InfAP VIREO	Mean InfAP VERGE	Mean InfAP ITEC_UU	Mean InfAP IMOTION	Prec. VIREO	Prec. VERGE	Prec. ITEC_UU	Prec. IMOTION
505	0.0006	0.0034	0.0006	0.0186	0.1875	0.6102	0.4103	0.4286
511	0.0058	0.0002	0.0027	0.0046	0.8500	0.4167	0.8235	0.6000
515	0.0048	0.0047	0.0007	0.0023	0.7143	0.5000	0.6667	0.7692
520	0.1182	0.1914	0.0003	0.0117	1.0000	0.7561	0.9487	0.9565
526	0.0152	0.0115	0.0002	0.0140	0.8958	0.9400	0.9149	0.7561

TABLE VII
SCORES OF COMMON TEAMS AT TRECVID AVS (INFAP ACROSS ALL RUNS) AND VBS (PRECISION). THE RESULTS ARE BASED ON THE EXTENDED GROUND TRUTH, INCLUDING THE ITEMS JUDGED DURING VBS.

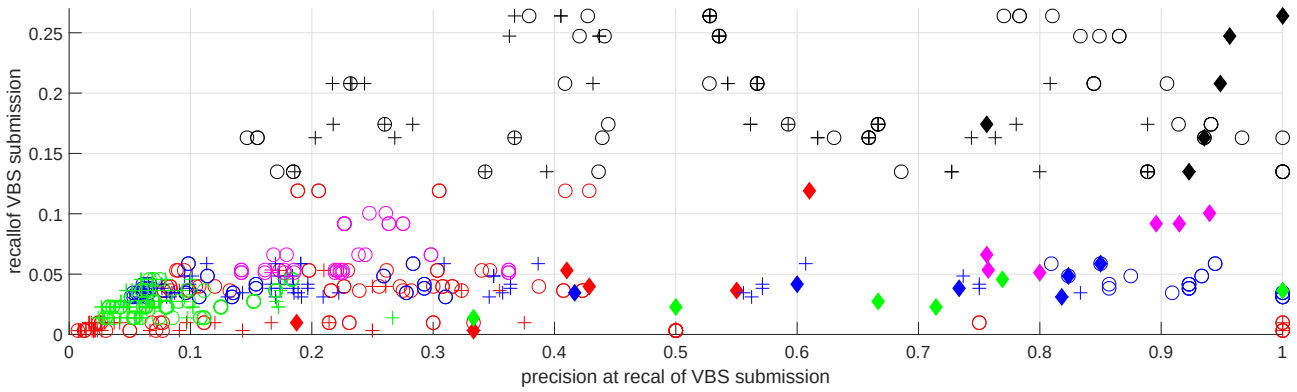

Fig. 7. Precision at the recall of the VBS submission (considering only the best TRECVID AVS submissions, where the desired recall level is reached within the top ranked 200 shots which are always judged). Diamonds indicate VBS submissions, plus signs fully automatic TRECVID AVS submissions and circles manually assisted TRECVID AVS submissions, the colors encode the task (red=505, blue=511, green=515, black=520, magenta=526). The results are based on the extended ground truth, including the items judged during VBS.

a detailed photographic memory is not usual in the population and so it should not be possible to exploit more and more specific details of the scene for filtering once the first attempts fail. In order to simulate a limited ability to remember details, the played scene should gradually get blurred.

In the textual KIS tasks, the text could simulate a potential conversation between a video retrieval tool expert and a witnessing user who is able to recognize the target scene. With the time running, the presented text description could be extended with additional facts, which simulates a discussion after the set of unsuccessful searches. Providing additional information could also help with the high complexity of the task. Note that four out of seven textual KIS tasks were not solved at VBS 2017 given the five minute time limit per task.

The AVS initialization will remain the same as at VBS 2017.

B. Evaluation objectives

Similar as in the previous period 2015-2017, the VBS main objective will be (revisited) evaluation scenarios for unconstrained visual/textual KIS and AVS tasks. The VBS will focus also on more complete comparability of tools and a stronger engagement of novice users.

According to our observations, textual KIS tasks are already highly challenging given a short task time limit for the 600h video collection. Although all six teams used high-quality video retrieval models, more than half of textual tasks were not solved. In such case, a question arises – what purpose is served to perform live evaluation with a too strict time limit? As mentioned in the previous section, an additional textual description could be gradually provided. As ”one image/scene is worth one thousand words”, adding a few more sentences to a task description does not have to be enough for solving the task by competing tools. As a new preference, we suggest splitting the VBS evaluation across two days (if possible). The

first day, more difficult (and also less entertaining) textual tasks will be evaluated separately without audience. Hence, a longer time limit could be allocated just for textual KIS tasks. The second day, the more entertaining part of VBS will take place, in front of the audience with easier visual KIS and AVS search tasks (a higher chance to find at least one scene as reported in Figure 4). The time frame for the second day will be more strict and so shorter time limits will be considered. Note that the search is exhausting and performing a lot of tasks during a long time period in one day is not feasible.

The VBS objectives should also focus more on the evaluation of the usability gap of the tools. We plan to fully re-introduce novice sessions for visual KIS and AVS tasks, where randomly selected users from the audience will control the tools without any help. In order to control for the effect of different skills, each tool will be controlled by a randomly selected team (one novice operator and novice assistants). First the experts will compete in several visual KIS and AVS tasks, while the assigned team of novice users will observe the tool interface and common usage of the tool. Then, for the second part of the evaluation only the novice users will try to solve the tasks. In case there are not enough volunteers, the tools with significantly low scores after all expert sessions could continue with expert users (without chance to beat a tool controlled by novice users). Note that use of novices also controls for using tasks previously used in TRECVID since TRECVID participants will not be selected from the audience.

C. Evaluation metrics

As presented in Section IV, the top performing tools at VBS influence to some extent the trends in the evolution of interactive video retrieval tools. As such, the VBS overall score evaluation methodology must be interpretable and fair. Based on six years of experience with the VBS competition, we suggest the following revisions to the most problematic evaluation metric concepts. These include the time limit, wrong submission penalization, the score function for AVS tasks and task category aggregation approaches.

The time limit $t_L \in R^+$ was introduced for practical reasons in order to make the live evaluation feasible with approximately 20 evaluated search sessions. At the same time, a small time limit goes against another goal – to fairly compare all the tools. A possible intuitive interpretation of the time limit could be a VBS preference in tools at least $\sum_{\forall video} video.length/(2t_L)$ times faster than the average task solve time on a classical sequential video player. Hence, tools slower than t_L receive zero points. According to our experience, in some cases the time limit causes a clearly unfair situation where one team finds the correct submission a few seconds before the time limit, while another team a few seconds after the time limit. Hence the tools performed almost the same, but the score reward was significantly different. Since it is cumbersome to assess such situations personally, we propose an approach (see Figure 8) providing a minimal time difference guarantee g between the last accepted correct submission at time $t_{last} \in R^+$ and the first potential late correct submission in task$_i$. The function $f^i_{KIS}(t, ws)$ can be updated to return zero only for $t^i_j > \max(t^i_L, t_{last}+g)$ or when it turns negative. Such an update also preserves the slope of linear $f^i_{TS}(t) = (t^i_L - t)/t^i_L$ in a given task category. The interpretation of t^i_L then becomes to be the initial time limit of a task$_i$, affecting the slope. Note that t_{last} can be updated repeatedly beyond t^i_L as we prefer potentially fewer search sessions with more fairly compared tools. As a consequence, the score gain can be smaller than c_S for solving a KIS task. On the other hand, c_S could be set higher to prefer tools with a higher number of solved KIS tasks with the introduced guarantee.

The second revisited concept is the wrong submission penalization. Its presence in the denominator significantly decreases the score after the third wrong submission, while the penalizing effect of a higher number of wrong submissions is lower. We propose a subtraction based penalization and a limited number of allowed submissions instead.

$$f^i_{KIS}(t, ws) = \lceil \max(0, s_C + (100 - s_C) \cdot f^i_{TS}(t) - f_{WSP}(ws)) \rceil$$

We believe that the subtraction based penalization better corresponds to a real world scenario, where more wrong submissions linearly increase time for their verification (e.g., by a witness). As $f_{WSP}(ws)$, we consider a simple linear function $s_P \cdot ws$, without free wrong submissions. s_P represents a score cost of checking one wrong submission and has a fixed time interval equivalent in the employed linear time score model (delay caused by checking). As f^i_{KIS} is supposed to be a non-negative function, the number of wrong submissions affecting the score is limited.

The new AVS score function represents the third significant change. We prefer to model the new score function as a precision-recall based measure, where teams are rewarded for each correct submission. The set of all correct shots in task$_i$ (for 100% recall estimation) can be obtained from the pooling of all correct submissions C^i_j from all k teams to a set $P^i = \bigcup_{j=1}^{k} C^i_j$. The new measure is designed to consider distinct videos and decrease the losses for not finding a video with a lot of correct shots. The suggested score function for team$_j$ in task$_i$ is defined for the pool P^i, the set of correct C^i_j and incorrect I^i_j submissions as:

$$f^i_{AVS}(C, I, P) = \left\lceil \frac{100 \cdot |C|}{|C| + |I|/2} \cdot \frac{|q(C)|}{|q(P)|} \right\rceil,$$

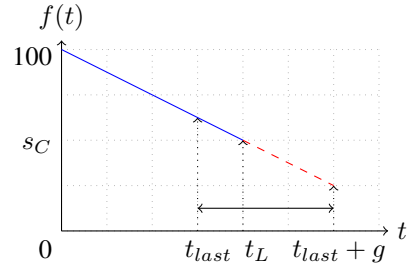

Fig. 8. A function $f(t)$ used to assign a score for each correct KIS task submission in a dynamically detected time limit.

where q is a "shot quantization" function merging temporarily close correct submissions. For each video separately, q considers a set of ranges obtained by a clustering of video shots present in the pool P^i. The function returns the set of distinct disjoint ranges intersecting the set of frame numbers f_j^i taken from a list of correct submissions (C or P). As a consequence, the function decreases the recall effect of videos with too many temporally close shots that are counted as correct. The number of wrong submissions is divided by 2 to decrease the effect of ambiguous scene descriptions and judgements. In order to motivate for collecting all correct shots in each range (comparability with TRECVID), we keep the original sets C_j^i, I_j^i in the first fraction. Hence, the teams can improve their precision by neglecting the effect of potentially non-empty I_j^i. If the users specify also their confidence about the submission (e.g., $\{sure, not\_sure\}$), the score function could be extended to combine the submission sets corresponding to different confidence levels. Note that for correct submissions with different confidence levels within the same range, the extended version could consider fractions of one recall point corresponding to the range.

The last revisited concept is the aggregation of the results of different task categories. Since there is no clear winner over aggregation strategies, we consider this change to be a new VBS preference for more universal tools. According to our observations, the score gains are not comparable among different categories (e.g., the higher complexity of textual KIS tasks leads to higher correct submissions times and thus lower scores in general). Instead of naively adding scores of each category together, scores will be max normalized and scaled to interval [0, 100] in each category c, i.e., the winner of a task category will always get 100 points. The overall score is computed as the average of all five task category sub-scores (visual KIS, textual KIS, AVS and two novice sessions for visual KIS and AVS). With respect to the score of the best team in a given category c (assumed to be non-zero), each of the k teams receive the following score:

$$score_{team_j} = \left\lceil \frac{1}{5} \cdot \sum_{c=1}^{5} \frac{100 \cdot score^c_{team_j}}{\max_{j=1..k}(score^c_{team_j})} \right\rceil .$$

D. Interaction logging

So far, the VBS competition reports consider only submission times and the number of correct/incorrect submissions logged by the VBS server. The relation between employed tool models/interfaces and the score is absent or provided only in an analysis by the team, often only based on the memory of the members. While we asked the teams to log complex interaction scenarios using a markup language schema and send them after the competition, only a small fraction of the teams provided the logs.

Based on this experience, we have decided to change the strategy for VBS events from 2018 on. First, the interaction logs will be a mandatory part of each submission. Hence, the VBS server will collect the logs during the competition. Second, as the tools use diverse models and interfaces, we plan to ask the teams to log only a generalized mandatory set of actions enabling the discovery of the main aspects of the search. For example, the search initialized with a color sketch and a keyword query followed by browsing several pages could be easily logged as `color; keyword; next page; ...; next page`. We hope that the simplicity of mandatory logging reduces necessary implementation efforts and will thus result in a more successful collection of the logs in the future. At the same time, such simple logs are sufficient to identify query initialization, search strategies and the frequency of browsing. In addition, each team can provide additional attributes (e.g., action time) and tool specific parameters in an optional part of the log (e.g., in brackets after each action).

According to our very recent experience at VBS 2018, eight out of nine teams implemented the simplified logging and submitted the logs to the VBS server.

VII. Future trends and open challenges in interactive video retrieval

Given the setting of VBS tasks constrained by a strict time limit, we can observe common features and trends of the successful tools at VBS. Their features include a frame-based data layer, effective query initialization by keywords/colors, sorted list filtering, grid-based (potentially sorted) organization of best matching keyframes, selected video inspection and browsing operations based on a similarity model. Automatic keyword annotation as well as similarity between two frames are mostly based on deep convolutional neural networks. Whereas for the visual KIS tasks the implemented features seem to be satisfactory to initiate the search with a good query and browse the results, for textual KIS tasks the systems still often fail. The time constraint, too many scenes corresponding to the actual query and limited information about the searched scene represent open challenges for future VBS events. Based on our observations and experience, in the following we present a list of challenges that should be addressed for more effective known-item and ad-hoc search.

More accurate frame annotation. The automatic image annotation methods based on state-of-the-art deep convolutional neural networks and transfer learning proved to be a promising approach to effectively filter the dataset and start browsing. With higher accuracy, the size of candidate sets for inspection can be further significantly decreased. The accurate identification of multiple objects and their relations in complex scenes will further improve the filtering power of the text-based queries combined with boolean operators and object localization.

Scene-based annotation and descriptors. Given a simple scene, the temporal axis becomes an important source of discriminative features. Probably due to a limited accuracy of the annotation methods for general scenes, the teams at VBS have not successfully employed scene-based annotations so far. However, once a critical level of accuracy is achieved, the scene-based annotations probably will be dominant for the query initialization phase. Effective scene based descriptors could help with similarity based browsing, once a good example scene is found.

Page zero problem. With an inappropriate query initialization, the options for browsing to a desired target are highly limited. Since the employed similarity models already consider activation features from deep neural networks, it would be interesting to employ a dynamic organization of frames into a set of semantically homogeneous clusters with selected representatives. An effective and at the same time efficient method would be welcome for refining larger candidate sets.

Diversity of visual representation. The analysis of textual KIS tasks reveals that users are often biased by their personal visualization of the scene, not considering the possible diversity in scene composition and object appearance. As many browsing tools were initially designed for retrieving visually similar content, they do not provide adequate support for exploring the potentially visually diverse set of keywords or concepts.

Mental target identification. By interacting with the system, users implicitly or even explicitly reveal valuable feedback. Based on the feedback, the system could adapt the underlying models to better fit the actual search needs and accelerate the retrieval process. One of the first attempts at VBS was based on a Bayesian framework, which proved to be a promising direction, considering the basic settings used by the tool. In future VBS events, it would be very interesting to observe more approaches applying state-of-the-art results from the relevance feedback and reinforcement learning areas.

More effective visualizations of results. According to our experience, users are often too focused on one particular memorized moment from the observed scene and overlook other moments from the scene that already appear in a simple grid-based result list visualization. From this perspective, suitable visualizations of the results (e.g., with a temporal context) could improve the effectiveness of the retrieval affected by the human factor. User studies addressing psychological aspects and subjectivity of the perception could provide better understanding of the limits of particular visualization methods.

VIII. Conclusion

The Video Browser Showdown is a well established live evaluation campaign, focusing on interactive retrieval in video archives. The results of the evaluation affect the evolution of participating tools, revealing promising trends in interactive video retrieval and provide insights into the complexity of the different task categories. Since 2017, the Video Browser Showdown has also included ad-hoc video search tasks evaluated at TRECVID. Hence, it is possible to evaluate and compare to some extent the performance of interactive and fully automatic video retrieval approaches. While relative task complexity is mostly comparable between the two benchmarks, the interactive tools are generally more successful in retrieving the most relevant items correctly, while being less successful in retrieving all relevant items due to the time limit t_L.

Based on the analysis of the data collected in three years of the competition, we revisited many aspects of Video Browser Showdown task design and scoring, in order to provide a more fair and realistic evaluation. For example, the considered evaluation metrics have been modified to better reflect the presented goals of the evaluation. We believe that all the suggested changes will foster progress in the evolution of interactive video retrieval tools, and make these tools more aligned with real-world challenges in applications of video retrieval.

Acknowledgments

This paper has been supported by Czech Science Foundation (GAČR) project Nr. 17-22224S. This work was also supported by Universität Klagenfurt and Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214 U. 3520/26336/38165.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Appendix
Simulated AP calculation

Given is an unranked result list with length n containing c relevant and $n-c$ non-relevant items. The total number of relevant items for the task is $nrel$ (with $c \leq nrel$). The aim is to determine an estimate of the AP for the list, determined as average of the AP over the permutations of the list.

As we have binary relevance, the list with length n and c relevant items has $\binom{n}{c}$ distinct permutations. Let L_p be the list of permutation p, then the mean AP is given as

$$\frac{1}{\binom{n}{c}} \sum_{p=1}^{\binom{n}{c}} \mathrm{AP}(L_p) = \frac{1}{\binom{n}{c}} \sum_{p=1}^{\binom{n}{c}} \frac{\sum_{k=1}^{n} \mathrm{rel}(L_p^k)\mathrm{Pr}(L_p^k)}{nrel} = \quad (2)$$

We expand the calculation of precision Pr() at rank k (the $k-$th position in the permutation):

$$\frac{1}{nrel\binom{n}{c}} \sum_{p=1}^{\binom{n}{c}} \sum_{k=1}^{n} \mathrm{rel}(L_p^k) \frac{1}{k} \sum_{j=1}^{k} \mathrm{rel}(L_p^j) = \quad (3)$$

We swap the two sums and move $\frac{1}{k}$ out:

$$\frac{1}{nrel\binom{n}{c}} \sum_{k=1}^{n} \frac{1}{k} \sum_{p=1}^{\binom{n}{c}} \mathrm{rel}(L_p^k) \sum_{j=1}^{k} \mathrm{rel}(L_p^j) \quad (4)$$

For a particular rank k, we can discard the case where $\mathrm{rel}(L_p^k) = 0$ (as the whole term evaluates to 0). For the remaining ranks, the inner sum counts the number of relevant elements up to the current rank, and the outer sums over the permutations of the list. We can replace this by a more efficient calculation of the number of relevant items per rank.

Determining the number of permutations that have a relevant item at rank k means fixing one relevant item in the list. Thus, the $c-1$ remaining returned relevant items can be arranged

at the remaining $n-1$ ranks, which means there are $\binom{n-1}{c-1}$ permutations that have a relevant item at rank k.

For each of the ranks $t = 1 \dots k-1$ above in the list, we consider the item at this rank to be relevant. This means that all items except for ranks k and t can be varied, which means that $c-2$ returned relevant items can be arranged on $n-2$ positions. Thus there are $\binom{n-2}{c-2}$ permutations with a relevant item at each rank t. The sum of relevant items up to rank k is thus given as

$$sumrel_k = \begin{cases} c, \text{if } c < 2, \\ \binom{n-1}{c-1} + (k-1)\binom{n-2}{c-2} \text{ otherwise,} \end{cases} \quad (5)$$

where we consider the cases of no relevant returned items ($c = 0$, obviously leading to 0 relevant items up to rank k) and a single relevant item ($c = 1$, with just a single relevant item up to rank k).

If we insert this into the equation above we obtain

$$\frac{1}{nrel\binom{n}{c}} \left(\sum_{k=1}^{n} \frac{sumrel_k}{k} \right). \quad (6)$$

Note that the runtime complexity of this method is $O(2n \cdot \min(c, n-c))$ due to performing n calculations of two binomial coefficients each time. This is significantly faster than the naive approach of evaluating average precision (complexity $O(n)$) for each permutation, i.e., $O(\binom{n}{c} \cdot n)$ in total.

Our evaluations on the TRECVID AVS submissions have shown that while simulated AP slightly underestimates the true AP for most submissions, using only the mean of the best and worst case of the list overestimates the true AP, with a significantly higher error.

REFERENCES

[1] G. Awad, A. Butt, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quénot, M. Eskevich, R. Ordelman, G. J. F. Jones, and B. Huet, "Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking," in *Proceedings of TRECVID 2017*. NIST, USA, 2017.

[2] M. Larson, M. Soleymani, G. Gravier, B. Ionescu, and G. J. Jones, "The benchmarking initiative for multimedia evaluation: Mediaeval 2016," *IEEE MultiMedia*, vol. 24, no. 1, pp. 93–96, 2017.

[3] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M. Naphade, A. Natsev, C. Neti, H. Nock *et al.*, "Ibm research trecvid-2003 video retrieval system," *NIST TRECVID-2003*, vol. 7, no. 8, p. 36, 2003.

[4] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 494–501.

[5] A. F. Smeaton, P. Wilkins, M. Worring, O. De Rooij, T.-S. Chua, and H. Luan, "Content-based video retrieval: Three example systems from trecvid," *International Journal of Imaging Systems and Technology*, vol. 18, no. 2-3, pp. 195–201, 2008.

[6] C. G. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, vol. 2, no. 4, pp. 215–322, 2008.

[7] M. Eskevich, R. Aly, D. Racca, R. Ordelman, S. Chen, and G. J. Jones, "The search and hyperlinking task at mediaeval 2014," 2014.

[8] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen, "The mediaeval 2015 affective impact of movies task." in *MediaEval*, 2015.

[9] O. De Rooij and M. Worring, "Browsing video along multiple threads," *IEEE Transactions on Multimedia*, vol. 12, no. 2, pp. 121–130, 2010.

[10] J. Yuan, Z.-J. Zha, Y.-T. Zheng, M. Wang, X. Zhou, and T.-S. Chua, "Utilizing related samples to enhance interactive concept-based video search," *IEEE Transactions on Multimedia*, vol. 13, no. 6, pp. 1343–1355, 2011.

[11] Y.-G. Jiang, J. Wang, Q. Wang, W. Liu, and C.-W. Ngo, "Hierarchical visualization of video search results for topic-based browsing," *IEEE Transactions on Multimedia*, vol. 18, no. 11, pp. 2161–2170, 2016.

[12] K. Schoeffmann, D. Ahlström, and M. A. Hudelist, "3-d interfaces to improve the performance of visual known-item search," *IEEE Transactions on Multimedia*, vol. 16, no. 7, pp. 1942–1951, 2014.

[13] K. Schöffmann and W. Bailer, "Video browser showdown," *SIGMultimedia Rec.*, vol. 4, no. 2, pp. 1–2, Jul. 2012. [Online]. Available: http://doi.acm.org/10.1145/2350204.2350205

[14] M. G. Christel, C. Huang, N. Moraveji, and N. Papernick, "Exploiting multiple modalities for interactive video retrieval," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 3. IEEE, 2004, pp. iii–1032.

[15] G. Gaughan, A. F. Smeaton, C. Gurrin, H. Lee, and K. McDonald, "Design, implementation and testing of an interactive video retrieval system," in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*. ACM, 2003, pp. 23–30.

[16] M. G. Christel and R. Yan, "Merging storyboard strategies and automatic retrieval for improving interactive video search," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, ser. CIVR '07. New York, NY, USA: ACM, 2007, pp. 486–493. [Online]. Available: http://doi.acm.org/10.1145/1282280.1282351

[17] C. G. Snoek, M. Worring, D. C. Koelma, and A. W. Smeulders, "A learned lexicon-driven paradigm for interactive video retrieval," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 280–292, 2007.

[18] K. Schoeffmann, M. A. Hudelist, and J. Huber, "Video interaction tools: A survey of recent work," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 14:1–14:34, Sep. 2015.

[19] M. Worring, P. Sajda, S. Santini, D. A. Shamma, A. F. Smeaton, and Q. Yang, "Where is the user in multimedia retrieval?" *IEEE MultiMedia*, vol. 19, no. 4, pp. 6–10, 2012.

[20] K. Schoeffmann, M. J. Primus, B. Muenzer, S. Petscharnig, C. Karisch, Q. Xu, and W. Huerst, *Collaborative Feature Maps for Interactive Video Search*. Cham: Springer International Publishing, 2017, pp. 457–462.

[21] K. Schoeffmann and F. Hopfgartner, "Interactive video search," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: ACM, 2015, pp. 1321–1322.

[22] K. Schoeffmann, "A user-centric media retrieval competition: The video browser showdown 2012-2014," *IEEE MultiMedia*, vol. 21, no. 4, pp. 8–13, 2014.

[23] P. Geetha and V. Narayanan, "A survey of content-based video retrieval," *Journal of Computer Science*, vol. 4, no. 6, p. 474, 2008.

[24] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 41, no. 6, pp. 797–819, Nov 2011.

[25] E. Yilmaz and J. A. Aslam, "Estimating average precision when judgments are incomplete," *Knowledge and Information Systems*, vol. 16, no. 2, pp. 173–211, Aug 2008.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[27] Y. Yan, S. Pouyanfar, Y. Tao, H. Tian, M. P. Reyes, M.-L. Shyu, S.-C. Chen, W. Chen, T. Chen, and J. Chen, "Florida international university-university of miami trecvid 2017," 2017.

[28] K. Ueki, K. Hirakawa, K. Kikuchi, T. Ogawa, and T. Kobayashi, "Waseda meisei at trecvid 2017: Ad-hoc video search."

[29] D. Francis, B. Merialdo, and B. Huet, "Eurecom at trecvid 2017: The adhoc video search."

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[31] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 675–678. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654889

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec 2015.

[33] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, 2017.

[34] K. Schoeffmann, M. Taschwer, and L. Boeszoermenyi, "The video explorer: A tool for navigation and searching within a single video based on fast content analysis," in *Proceedings of the First Annual ACM SIGMM Conference on Multimedia Systems*, ser. MMSys '10. New York, NY, USA: ACM, 2010, pp. 247–258. [Online]. Available: http://doi.acm.org/10.1145/1730836.1730867

[35] O. de Rooij, J. van Wijk, and M. Worring, "Mediatable: Interactive categorization of multimedia collections," *Computer Graphics and Applications, IEEE*, vol. 30, no. 5, pp. 42–51, Sept 2010.

[36] H. Luan, Y.-T. Zheng, M. Wang, and T.-S. Chua, "Visiongo: Towards video retrieval with joint exploration of human and computer," *Information Sciences*, vol. 181, no. 19, pp. 4197 – 4213, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025511002672

[37] B. Adams, S. Greenhill, and S. Venkatesh, "Towards a video browser for the digital native," in *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, July 2012, pp. 127–132.

[38] M. Ferecatu and D. Geman, "A statistical framework for image category search from a mental picture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 1087–1101, 2009.

[39] A. G. Hauptmann, W.-H. Lin, R. Yan, J. Yang, and M.-Y. Chen, "Extreme video retrieval: Joint maximization of human and computer performance," in *Proceedings of the 14th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '06. New York, NY, USA: ACM, 2006, pp. 385–394. [Online]. Available: http://doi.acm.org/10.1145/1180639.1180721

[40] W. Hürst, R. van de Werken, and M. Hoet, *A Storyboard-Based Interface for Mobile Video Browsing*. Cham: Springer International Publishing, 2015, pp. 261–265.

[41] W. Hürst, A. I. V. Ching, K. Schoeffmann, and M. J. Primus, "Storyboard-based video browsing using color and concept indices," in *International Conference on Multimedia Modeling*. Springer, 2017, pp. 480–485.

[42] D. A. Keim, "Information visualization and visual data mining," *IEEE Trans. Vis. Comput. Graph.*, vol. 8, no. 1, pp. 1–8, 2002.

[43] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. P. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1604–1613, 2014.

[44] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, "What you see is what you can change: Human-centered machine learning by interactive visualization," *Neurocomputing*, vol. 268, pp. 164–175, 2017.

[45] J. Lokoč, P. A. Nguyen, M. Vomlelová, and C. Ngo, "Color-sketch simulator: A guide for color-based visual known-item search," in *Advanced Data Mining and Applications - 13th International Conference, ADMA 2017, Singapore, November 5-6, 2017, Proceedings*, 2017, pp. 754–763.

[46] W. Bailer and H. Rehatschek, "Comparing fact finding tasks and user survey for evaluating a video browsing tool," in *Proceedings of ACM Multimedia*, Beijing, CN, Oct. 2009.

[47] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA: ACM Press, 2006, pp. 321–330.

[48] G. Awad, W. Kraaij, P. Over, and S. Satoh, "Instance search retrospective with focus on trecvid," *International Journal of Multimedia Information Retrieval*, vol. 6, no. 1, pp. 1–29, 2017.

[49] G. Awad, C. G. M. Snoek, A. F. Smeaton, and G. Qunot, "Trecvid semantic indexing of video: A 6-year retrospective," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 3, pp. 187–208, 2016, invited paper.

[50] C. G. Snoek, M. Worring, O. de Rooij, K. E. van de Sande, R. Yan, and A. G. Hauptmann, "Videolympics: Real-time evaluation of multimedia retrieval systems," *IEEE MultiMedia*, vol. 15, no. 1, 2008.

[51] *CIVR '07: Proceedings of the 6th ACM International Conference on Image and Video Retrieval*. New York, NY, USA: ACM, 2007.

[52] K. Schoeffmann, D. Ahlström, W. Bailer, C. Cobârzan, F. Hopfgartner, K. McGuinness, C. Gurrin, C. Frisson, D.-D. Le, M. Del Fabro *et al.*, "The video browser showdown: a live evaluation of interactive video search tools," *International Journal of Multimedia Information Retrieval*, vol. 3, no. 2, pp. 113–127, 2014.

[53] A. Blažek, J. Lokoč, F. Matzner, and T. Skopal, *Enhanced Signature-Based Video Browser*. Cham: Springer International Publishing, 2015, pp. 243–248.

[54] K. U. Barthel, N. Hezel, and R. Mackowiak, *Navigating a Graph of Scenes for Exploring Large Video Collections*. Cham: Springer International Publishing, 2016, pp. 418–423.

[55] L. Rossetto, I. Giangreco, C. Tănase, H. Schuldt, S. Dupont, and O. Seddati, *Enhanced Retrieval and Browsing in the IMOTION System*. Cham: Springer International Publishing, 2017, pp. 469–474.

[56] L. Rossetto, I. Giangreco, H. Schuldt, S. Dupont, O. Seddati, M. Sezgin, and Y. Sahillioğlu, *IMOTION — A Content-Based Video Retrieval Engine*. Cham: Springer International Publishing, 2015, pp. 255–260.

[57] M. A. Hudelist, C. Cobârzan, C. Beecks, R. van de Werken, S. Kletz, W. Hürst, and K. Schoeffmann, *Collaborative Video Search Combining Video Retrieval with Human-Based Visual Inspection*. Cham: Springer International Publishing, 2016, pp. 400–405.

[58] D. Kuboň, A. Blažek, J. Lokoč, and T. Skopal, *Multi-sketch Semantic Video Browser*. Cham: Springer International Publishing, 2016, pp. 406–411.

[59] A. Blažek, J. Lokoč, and D. Kuboň, *Video Hunter at VBS 2017*. Cham: Springer International Publishing, 2017, pp. 493–498.

[60] K. U. Barthel, N. Hezel, and R. Mackowiak, *Graph-Based Browsing for Large Video Collections*. Cham: Springer International Publishing, 2015, pp. 237–242.

[61] A. Moumtzidou, T. Mironidis, E. Apostolidis, F. Markatopoulou, A. Ioannidou, I. Gialampoukidis, K. Avgerinakis, S. Vrochidis, V. Mezaris, I. Kompatsiaris, and I. Patras, *VERGE: A Multimodal Interactive Search Engine for Video Browsing and Retrieval*. Cham: Springer International Publishing, 2016, pp. 394–399.

[62] Y.-J. Lu, P. A. Nguyen, H. Zhang, and C.-W. Ngo, *Concept-Based Interactive Search System*. Cham: Springer International Publishing, 2017, pp. 463–468.

[63] T. D. Ngo, V.-T. Nguyen, V. H. Nguyen, D.-D. Le, D. A. Duong, and S. Satoh, *NII-UIT Browser: A Multimodal Video Search System*. Cham: Springer International Publishing, 2015, pp. 278–281.

[64] J. He, X. Shang, H. Zhang, and T.-S. Chua, *Mental Visual Browsing*. Cham: Springer International Publishing, 2016, pp. 424–428.

[65] V.-T. Nguyen, T. D. Ngo, D.-D. Le, M.-T. Tran, D. A. Duong, and S. Satoh, *Semantic Extraction and Object Proposal for Video Search*. Cham: Springer International Publishing, 2017, pp. 475–479.

[66] A. Moumtzidou, K. Avgerinakis, E. Apostolidis, F. Markatopoulou, K. Apostolidis, T. Mironidis, S. Vrochidis, V. Mezaris, I. Kompatsiaris, and I. Patras, *VERGE: A Multimodal Interactive Video Search Engine*. Cham: Springer International Publishing, 2015, pp. 249–254.

[67] L. Rossetto, I. Giangreco, S. Heller, C. Tănase, H. Schuldt, S. Dupont, O. Seddati, M. Sezgin, O. C. Altıok, and Y. Sahillioğlu, *IMOTION – Searching for Video Sequences Using Multi-Shot Sketch Queries*. Cham: Springer International Publishing, 2016, pp. 377–382.

[68] A. Moumtzidou, T. Mironidis, F. Markatopoulou, S. Andreadis, I. Gialampoukidis, D. Galanopoulos, A. Ioannidou, S. Vrochidis, V. Mezaris, I. Kompatsiaris, and I. Patras, "VERGE in VBS 2017," in *MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part II*. Cham: Springer International Publishing, 2017, pp. 486–492.

[69] W. Bailer, W. Weiss, and S. Wechtitsch, *Selecting User Generated Content for Use in Media Productions*. Cham: Springer International Publishing, 2016, pp. 388–393.

[70] L. Rossetto, I. Giangreco, C. Tănase, H. Schuldt, S. Dupont, O. Seddati, M. Sezgin, and Y. Sahillioğlu, *iAutoMotion – an Autonomous Content-Based Video Retrieval Engine*. Cham: Springer International Publishing, 2016, pp. 383–387.

[71] J. Lokoč, A. Blažek, and T. Skopal, "Signature-based video browser," in *Proceedings of the 20th Anniversary International Conference on MultiMedia Modeling - Volume 8326*, ser. MMM 2014. New York, NY, USA: Springer-Verlag New York, Inc., 2014, pp. 415–418.

[72] I. Giangreco and H. Schuldt, "ADAM pro: Database support for big multimedia retrieval," *Datenbank-Spektrum*, vol. 16, no. 1, pp. 17–26, 2016.

[73] C. Beecks, S. Kirchhoff, and T. Seidl, "On stability of signature-based similarity measures for content-based image retrieval," *Multimedia Tools and Applications*, pp. 1–14, 2013.

[74] J. Lokoc, A. Blažek, and T. Skopal, "On effective known item video search using feature signatures," in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 524.

[75] C. Beecks, S. Kirchhoff, and T. Seidl, "Signature matching distance for content-based image retrieval," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 2013, pp. 41–48.