# A Task Category Space for User-Centric Comparative Multimedia Search Evaluations

Jakub Lokoč[1 [0000−0002−3558−4144]], Werner Bailer[2 [0000−0003−2442−4900]], Kai Uwe Barthel[3 [0000−0001−6309−572X]], Cathal Gurrin[4 [0000−0003−2903−3968]], Silvan Heller[5 [0000−0001−5386−330X]], Björn Þór Jónsson[6 [0000−0003−0889−3491]], Ladislav Peška[1 [0000−0001−8082−4509]], Luca Rossetto[7 [0000−0002−5389−9465]], Klaus Schoeffmann[8 [0000−0002−9218−1704]], Lucia Vadicamo[9 [0000−0001−7182−7038]], Stefanos Vrochidis[10 [0000−0002−2505−9178]], and Jiaxin Wu[11 [0000−0003−4074−3442]]

[1] Charles University, Prague, Czech Republic
jakub.lokoc@matfyz.cuni.cz, ladislav.peska@matfyz.cuni.cz
[2] JOANNEUM RESEARCH, Graz, Austria werner.bailer@joanneum.at
[3] HTW Berlin, Berlin, Germany barthel@htw-berlin.de
[4] Dublin City University, Dublin, Ireland cathal.gurrin@dcu.ie
[5] University of Basel, Basel, Switzerland silvan.heller@unibas.ch
[6] IT University of Copenhagen, Copenhagen, Denmark bjth@itu.dk
[7] University of Zurich, Zurich, Switzerland rossetto@ifi.uzh.ch
[8] Klagenfurt University, Klagenfurt, Austria ks@itec.aau.at
[9] ISTI CNR, Pisa, Italy lucia.vadicamo@isti.cnr.it
[10] Centre for Research and Technology Hellas, Thessaloniki, Greece stefanos@iti.gr
[11] City University of Hong Kong, Hong Kong, China jiaxin.wu@my.cityu.edu.hk

**Abstract.** In the last decade, user-centric video search competitions have facilitated the evolution of interactive video search systems. So far, these competitions focused on a small number of search task categories, with few attempts to change task category configurations. Based on our extensive experience with interactive video search contests, we have analyzed the spectrum of possible task categories and propose a list of individual axes that define a large space of possible task categories. Using this concept of category space, new user-centric video search competitions can be designed to benchmark video search systems from different perspectives. We further analyse the three task categories considered so far at the Video Browser Showdown and discuss possible (but sometimes challenging) shifts within the task category space.

**Keywords:** Multimedia Retrieval · Task Taxonomy · Evaluation Design

## 1 Introduction

The explosion in the production and diffusion of multimedia data over the last decades has triggered a strong interest toward the development of systems for

This is a pre-copyedit version of this article. The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-98358-1_16

the storage, management and retrieval of large-scale multimedia archives. While Information Retrieval (IR) [26] approaches initially focused mainly on text documents, since the 1990s there has been flourishing research activity on content-based retrieval systems for other types of media, such as images and videos [9,13]. More recently, the synergy between IR and Artificial Intelligence techniques has enabled the development of retrieval systems that also support cross-modal searches and multiple query types (e.g., [12,17,20,35,36]). Furthermore, the suggestion that information retrieval is more appropriately regarded as an inherently interactive/evolving process [4,5] paved the way for the design of interactive and more user-centric systems, where the query expressing the user's information need is no longer considered as predetermined and static, but rather evolves dynamically during a search process [7,19,27,28,34].

To evaluate the effectiveness of different video retrieval systems, some benchmarking competitions have been established [8,11,28,29,31]. For example, VideOlympics [31] conducted assessments of interactive video retrieval systems on Ad-hoc Video Search (AVS) in 2007–2009, and the Video Browser Showdown (VBS) [28] has started to assess visual Known-Item Search (KIS) tasks since 2012. However, existing competitions focus on a small number of task categories, e.g., VBS only evaluates AVS, visual KIS, and textual KIS tasks. Moreover, the task design space is not well understood.

The main contribution of this paper is a structured description of a task category space for user-centric video search competitions. While existing evaluation initiatives are recapitulated in the related work section, a comprehensive task category space, based on our long-term experience with interactive video search competitions, is presented in Section 3. Three popular task categories are revisited in Section 4, perceiving the tasks as elements of the proposed space and discussing possible future options for designing a rich set of benchmark activities.

## 2 Related Work and Background

Different interactive multimedia retrieval systems naturally have distinct user interfaces, browsing, and searching functionalities, which introduce a bias in users' attitudes when formulating and refining their queries. Hence, the comparative performance of interactive retrieval systems cannot be easily evaluated and compared outside of controlled environments, set up for this specific purpose, such as benchmarking campaigns.

In this context, interactive competitions such as the Video Browser Showdown (VBS) [19,24,28] and the Lifelog Search Challenge (LSC) [10,11] provide an equitable performance assessment of interactive retrieval systems, where not only the same search tasks on the same dataset are employed, but also users with different level of knowledge of systems (i.e., expert and novice users) are involved in the evaluation process in a live real-time benchmarking activity. The VBS video search competition comprises three search tasks, namely Ad-hoc Video Search (AVS), visual Known-Item Search (KIS) and textual KIS. Textual KIS tasks are also evaluated in LSC but for the use case of multimodal lifelog data

retrieval. Although not focusing specifically on the case of interactive systems, other evaluation campaigns have played an important role in the assessment of multimedia retrieval and analysis techniques for a wide variety of tasks (see Table 1). For example, the Benchmarking Initiative for Multimedia Evaluation (MediaEval) [8,15] has offered a large spectrum of tasks (almost 50 different tasks since 2010) related to multimedia retrieval, analysis, and exploration with a focus on the human and social aspects of multimedia. The TREC Video Retrieval Evaluation (TRECVID) [1,30] over the last two decades has spawned over twenty tasks related to content-based analysis of and retrieval from digital video, including automatic AVS, KIS, Video Hyperlinking, and the automatic detection of a variety of semantic and low-level video features [1,2,22,29]. In the context of video understanding and search, in 2020 the Video Pentathlon Workshop[1] offered an interesting challenge that tackles the task of caption-to-video retrieval, i.e. retrieving relevant videos using only natural language queries.

The task of retrieving one particular data item that satisfies a very specific information need for a user (i.e., KIS task) recurs in many benchmarking campaigns. Although there is no universally accepted definition of KIS in the multimedia domain, this concept originates in the field of library science, where it refers to the task of locating and obtaining a particular book or document of a catalogue that the searcher has in mind (e.g., the searcher knows the author, the title, or other distinguishing characteristics) [18,33]. Traditionally, there was a distinction between the concept of "Known-Item Search" (understood as the search for a particular document of which the details are known) and that of "subject search" (where the need is to locate material dealing with a particular subject or to answer a particular question) [32,6]. Walker and Janes [32] argued that the subject search is far more challenging than Known-Item Search, since it focuses on searching for what is not known, or perhaps does not exist, whereas bibliographic utilities can be employed to easily find a specific document (e.g. using the ISBN). However, arguments of this kind lose significance when the data to be handled is unstructured; in fact, searching for a specific multimedia item, say for example an image, without having a copy of the digital item at hand is hardly easier than searching for a generic image within a given topic. Lee et al. [16] made an important step toward a generalization of the KIS definition used in library and information science that could be transferred to multimedia search. They reviewed "conceptual" understandings of known-item search (e.g., looking for something the user knows exists) that are independent of the "operational" definitions designed to find the item of interest in a particular context (e.g., search a card catalogue for an item using the author or the title). In the 2010 edition of TRECVID, the KIS task category was formulated in the video search domain as one that "models the situation in which someone knows of a video, has seen it before, believes it is contained in a collection, but doesn't know where to look" [21]. However, in [21] it is also assumed that to begin a search, the user formulates a text-only description which captures what they remember about the target video. This "operational" constraint, however, is too restric-

[1] https://www.robots.ox.ac.uk/~vgg/challenges/video-pentathlon/challenge.html

**Table 1.** An overview of prominent benchmarking campaigns with multimedia retrieval and analysis tasks.

| **Benchmarking Campaign** | **Evaluated Tasks (years)**[1] |
|---|---|
| TRECVid (since 2001)[2] | *Ad-hoc Video Search* (2003-2009, 2016-2021)[3], *Instance search* (2010-2021), *Surveillance event detection* (2008-2017), *High-level Semantic Feature Extraction* (2002-2009), *Semantic indexing (and Localization)* (2010-2016), *Shot boundary detection* (2001-2007), *Multimedia Event Detection* (2010-2017), *Activities in Extended Video* (2018-2021), *Video to Text Description* (2017-2021), *Content-based multimedia copy detection* (2008-2011), *Streaming Multimedia Knowledge Base Population task* (2018, 2019, 2021), *Video Hyperlinking*[4] (2015-2017), *Multimedia event recounting* (2012-2014), *Textual Known-item search* (2010-2012), *Rushes Exploitation and Summarization* (2006-2008), *Video Summarization* (2020, 2021), *Disaster Scene Description and Indexing* (2020, 2021), *Story segmentation* (2003, 2004), *Social-Media Video Storytelling Linking* (2018), *Low-level Feature Extraction* (2005). |
| VideOlympics (2007-2009) | *Ad-hoc Video Search* (2007-2009) |
| MediaEval (since 2010) | *Emotional Impact of Movies (including Boredom and Violent Scenes Detection)* (2010-2018), *Emotions (and Themes) in Music* (2013-2015,2019-2021), *Multimodal geo-location prediction* (2010-2012,2014-2016), *Medico Multimedia* (2017-2021), *Retrieving Diverse Social Images* (2013-2017), *Predicting Media Memorability* (2018-2021), *C@merata: Querying Musical Scores* (2014-2017), *Query by Example Search on Speech* (2012-2015), *Social Event Detection* (2011-2014), *Insight for Wellbeing* (2019-2021), *Sports Video* (2019-2021), *Pixel Privacy* (2018-2020), *Multimedia Satellite* (2017-2019), *Video Search and Hyperlinking* (2012-2014), *Visual Privacy* (2012-2014), *Video Genre Tagging* (2010-2012), and other 32 tasks appearing in fewer than three editions. |
| VBS (since 2012) | *Visual Know-Item Search* (2012-2021), *Textual Know-Item Search* (2014-2021), *Ad-hoc Video Search* (2017-2021) |
| LSC (since 2018) | *Known-Item search* (multimodal lifelog data) (2018-2021) |
| Video Pentathlon (2020) | *Video Retrieval using natural language queries (text-to-video cross-modal retrieval)* (2020) |

[1] According to the information provided on the web pages of the respective evaluation campaigns: TRECVid (https://trecvid.nist.gov/), MediaEval (https://multimediaeval.github.io/), VBS (https://videobrowsershowdown.org/), LSC (http://lsc.dcu.ie), Video Pentathlon (https://www.robots.ox.ac.uk/~vgg/challenges/video-pentathlon/).

[2] Sarted as a video track featured in the 2001 and 2002 at the **T**ext **RE**trieval **C**onference and then became an independent evaluation campaign in 2003.

[3] Since 2016 in collaboration with VBS for the evaluation of interactive systems.

[4] Previously run in MediaEval.

tive as other search strategies may be employed to implement the Known-Item Search. For example, a user may have access to systems that support several search modes (e.g., search by sketch, search by color, etc.) that can be used alone or in combination with text-based queries.

The task of Ad-hoc Video Search (AVS) focuses on general search. The information need is formulated as a textual topic containing a person, action, object, location, etc., and a combination of them, such as *"Find shots of one or more people walking or bicycling on a bridge during daytime"* [3,14]. Given a topic, the participants need to return relevant video segments from the test collection which satisfy the need. Its history dates back to as early as 2003 when TRECVID established it as a video search task [14]. In the beginning, the task included humans in the loop to search relevant videos in manual or interactive settings (allowing users to reformulate the query). It also allows fully automatic submission since 2005. As the TRECVID AVS task was intractable, it was replaced by the Multimedia Event Detection from 2010-2017 to promote the progress of zero-example video event search [21]. However, in 2016 the AVS task has been resumed to promote a more realistic setting where not only events are used as retrieved topics [3]. While TRECVID AVS focuses on evaluating the effectiveness of the submitted search systems, VideOlympics [31] was established as a live competition from 2007 to 2010 to evaluate the influence of interaction behaviors and the visualized interface of the search systems on answering the AVS task, a similar goal to the one pursued by VBS. Since 2016, VBS has started to work with TRECVID to evaluate AVS in the interactive setting [27].

## 3 Task Category Space for User-Centric Video Search

Whereas related evaluation initiatives provide a large spectrum of task types for multimedia data (classification, analysis, prediction, retrieval, linking, etc.), this paper focuses on a list of options for interactive video search competitions. Nevertheless, we demonstrate that there are still many options to construct an interactive search task category and that only a negligible fraction of categories is currently addressed by evaluation campaigns.

Specifically, we present several domains describing a video search task category from different perspectives. In the following, each paragraph describes one *domain axis* $A_{d_i}$, constituting a particular aspect of content-based search evaluation. Their combination $A_{d_1} \times \ldots \times A_{d_n}$ forms the entire space of various task categories for user-centric video search evaluations. Elements of the space can then be selected to design a new user-centric video search campaign. Please note that there are combinations resulting in equivalent task types (e.g., for one target item in $A_{CI}$, all $A_{SI}$ options request the one item), and some combinations may not model actual real-world problems.

### $A_{CI}$: Number of correct items satisfying a search need in a dataset

1. One target item which is assumed to be unique (i.e., differences to near duplicates are considered relevant). Identical copies can be considered equivalent.

2. All near-duplicates of one target item, there is binary relevance since the near-duplicates depict the same content.
3. Many semantically related items, potentially with multi-valued relevance.

**$A_{SI}$: Requested number of submitted correct items**

1. One correct item is enough (e.g., searching for an evidence).
2. A limited number of correct items is sufficient, allowing variety of choices.
3. As many as possible correct items are requested, focusing on recall.

**$A_{PM}$: Search need presentation modality**

1. Visual and auditory experience (no recording is allowed, however, only human perception).
2. Provided text description.
3. Combined perception and text information.

**$A_{PT}$: Search need presentation timing**

1. Task specification is revealed a longer time (e.g., one hour) before the competition, some reminiscence clues may be presented during evaluation time.
2. All info about a task is revealed in the beginning of the task evaluation.
3. Search clues are gradually revealed during the task evaluation.

**$A_{PQ}$: Search need presentation quality**

1. Comprehensive presentation of search need (e.g., scene playback or exhaustive text description with all details present in the scene).
2. Presentation of limited information to solve a task (e.g., blurring or short abstract description with selected unique details).
3. Intentional introduction of unreliable or uncertain information.

**$A_{DC}$: Data collection**

1. Whole dataset is known in advance, teams search in the whole dataset.
2. Limited subset of the known dataset is specified when a task starts.
3. Completely new video is provided when task evaluation starts, fast online preprocessing is needed.

**$A_{TL}$: Time limit**

1. Limited time to solve a task, time limit controlled by an evaluation server.
2. Unrestricted time interval, teams search until they solve the task or give up (though the time for the whole competition can be limited).

**$A_{US}$: User skills**

1. Expert users who created the tool operate the system.
2. Novice users who are representative of a typical user without knowledge of how the system works and generates results operate the system.
3. Novice users who also have little experience operating computers in general operate the system.

**$A_{NU}$: Number of operating users**

1. One user should be able to solve a task. If there are more users per team, each user has to solve the task independently from the other team members.
2. Multiple users can cooperate to solve a task in a collaborative retrieval process.

**$A_{QM}$: Quality measure**

1. Time of submission, where faster systems are preferred.
2. Quantity of submitted relevant multimedia objects is important.
3. Relevance of the submitted objects is reflected, penalizing for incorrect submissions.
4. Diversity of submitted correct items is preferred.
5. Any meaningful combination of the first four options in $A_{QM}$ can be utilized as well. The combinations are not included due to space constraints.

The presented list comprises aspects that influence a task category. The last axis $A_{QM}$ enables to fix a competition evaluation preference, which affects the strategy that teams use to solve a task specified with the previous nine aspects. The main motivation for the task space was to analyze and present a high variety of options affecting task category design. Based on the presented list of axes and options, a large number of task categories can be identified. Whereas VBS currently uses just three of them, many other evaluation campaigns could be established for other interesting elements of the category space. For example, competitions focusing on online video preprocessing and search are definitely missing in the current pool of user-centric video search competitions.

## 4 Challenges of Task Categories Considered so far

In the following, the task types used in previous instances of the Video Browser Showdown are discussed in the context of the space presented in the previous section. Table 2 shows how these task types can be classified along the defined axes. The tasks represent only a small fraction of possible tasks in the space.

### 4.1 Visual KIS task

The Visual Known Item Search task presents a unique video sequence of roughly 20 seconds in length and asks for the exact sequence to be retrieved from the collection within at most 5 minutes. It rewards correct results which are produced quickly while penalizing incorrect results and long retrieval times.

**Advantages.** This relatively simple task setup has the advantage that it is easy to understand and has little ambiguity (assuming the target sequence is unique with respect to the collection). The resulting low barrier of entry makes this task type attractive for new teams or approaches, while at the same time encouraging the use of different query modes targeting different media modalities. Since there

**Table 2.** VBS'21 task categories represented as vectors of the task space. For each category, the value in an axis column presents the currently used axis option (specified in Section 3). Due to the virtual conference setting, only expert users were participating.

| **Task Name** | $A_{CI}$ | $A_{SI}$ | $A_{PM}$ | $A_{PT}$ | $A_{PQ}$ | $A_{DC}$ | $A_{TL}$ | $A_{US}$ | $A_{NU}$ | $A_{QM}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Visual KIS | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1, 3 |
| Textual KIS | 1 | 1 | 2 | 3 | 2 | 1 | 1 | 1 | 2 | 1, 3 |
| Ad-hoc search | 3 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2, 3 |

is no translation or transformation of any kind between the presented sequence and the expected result, the task does also not penalize certain teams over others as a side effect, for example based on language proficiency.

**Disadvantages.** The primary disadvantage of this task type is that it only roughly approximates a realistic scenario. In a situation where somebody tries to retrieve a previously observed video segment, attention and memory play a more substantial role than if the segment can be observed at the time of retrieval.

**Future options.** A possible change to make the task more accurately represent a realistic scenario would be to present a number of scenes some time before the competition, and then let users search for some of them ($A_{PT} : 2 \rightarrow 1$). This would ensure only the memorized aspects are available to them. Depending on the memory prompt, this could impact $A_{PM}$. However, in order to adjust the experiment for the different capabilities of humans to memorize visual information, this would require a large group of searchers who are not available at VBS. Another approach would be to explicitly model attention and memory effects and modify the presented sequence to simulate them ($A_{PQ} : 1 \rightarrow 2$). While there have been some early experiments in this direction [23], it is largely unclear how such effects could be simulated realistically and further research on query representation methods modeling human visual memory is needed.

### 4.2 Textual KIS task

The Textual Known Item Search task presents a textual description intended to uniquely describe a video sequence of roughly 20 seconds in length. The textual description becomes more detailed over time, often uniquely describing the intended sequence only in its most elaborate form. The task uses the same scoring mechanism as the previous task.

**Advantages.** This task models a realistic setting where a searcher has a limited recollection of a scene while having a clear search interest. It also works as a stand-in for a situation where the person with the search interest needs to verbally describe it to somebody else, who is then performing the search. Due to the inherent loss of detail in a textual description, when compared to the original sequence, the task also implicitly addresses inaccuracies in memory.

**Disadvantages.** The primary disadvantage of this task is its potential for ambiguity which can arise by several means. Since not all information is revealed at the start of the task, it is possible to submit sequences which match the currently available information but differ from the target sequence in a detail not yet known to participants. The limitations in textual descriptions, especially for non-native speakers, also make it difficult to establish a common understanding of the target. These problems would likely not occur to this degree in a real world setting, since there could exist a bidirectional communication channel between the person describing the search interest and the person performing the search.

**Future options.** To decrease ambiguity of text descriptions, they could be complemented with visual information, such as a hand-drawn sketch of some target frame, thus turning the task into text-induced KIS ($A_{PM} : 2 \rightarrow 3$). A sketch may provide a better understanding of the composition of a scene, based on memories of the task creator. At the same time, only limited and distorted information is added to the task description and thus the task remains challenging. Currently, users browse result sets based on their own imagination of a scene; adding visual hints would make users more efficient. Further, contextual information could be provided, such as events which occur in the same video but outside of the target sequence. Another option is provide the possibility to ask questions and clarify ambiguities during the task (a specific case of $A_{PT} = 3$). While this equalizes conditions for teams, it hinders reproducibility of the task.

### 4.3 Ad-hoc Video Search tasks

The Ad-hoc Video Search task provides a short textual description of a video topic. In contrast to the previously described tasks, this description is not unique to one sequence but can describe an arbitrary number of sequences. Since it is not feasible to annotate the entire dataset beforehand, submissions are assessed by human judges in a live manner.

**Advantages.** Several of the advantages of this task lie in its different nature compared to the previously discussed KIS tasks. In contrast to those, it models a less clearly specified search intent, where several results might satisfy an information need, which is also common in practice. The task also serves as a platform for more general text-to-video search approaches. Due to the potentially large number of relevant sequences, it also offers a non-binary outcome and lowers the burden for novel, experimental approaches.

**Disadvantages.** The disadvantages of this task come primarily from the need for human judges. Since the description of a task target is rather short, it is difficult to establish a common understanding on what constitutes a correct sequence. This can lead to misunderstandings between judges and participants as well as to inconsistent judgements if there is no clear understanding between judges. Another difficulty with this task type is that it is unclear how to best compare several sets of retrieved results and hence how to score them. An emphasis on pooled recall (i.e., recall established from correct submissions of all

teams) discourages competitors to share their methods or extracted data, while an emphasis on precision exaggerates the previously discussed problems with a shared understanding of the task.

**Future options.** The largest potential for improvement in this task lies arguably in the way correctness of retrieved results is assessed and how these assessments are aggregated into an overall score ($A_{QM} : 3$). To counteract possible inconsistencies in judgement, it would be possible to have multiple judges assess each retrieved sequence. Since the aggregation of multiple such assessments can not necessarily be losslessly presented in a binary format, it would be feasible to use multi-valued judgements directly. In order to better synchronize all judges and teams, clarifying questions could be allowed before the start of a task, or even during the task ($A_{PT} : 2 \rightarrow 3$). Another possible avenue to reduce possible confusion would be to augment the description with a series of example images depicting true positives and true negatives ($A_{PM} : 2 \rightarrow 3$). To avoid overlap with KIS tasks, these examples can be taken from outside the dataset. Similarly to Textual KIS, a paused task phase for establishing a common understanding on the scope of the query could be introduced.

## 5 Conclusions

This paper presented an overview of evaluation efforts in the video retrieval area and proposed a task category space covering many aspects of video search tasks. The space may inspire variations in tasks considered so far, as well as initiate novel campaigns focusing on currently missing user-centric benchmarks. We believe that in the future, interactive video search systems could be tested by multiple (even remote) evaluation campaigns using already established software tools [25]. The systems already participating at VBS and LSC, as well as many potential new systems, could prove their capabilities in a larger spectrum of tasks, ranging in various options listed in the presented task category space.

**Acknowledgments:** This work has been supported by Czech Science Foundation (GAČR) project 19-22071Y, by Science Foundation Ireland under grant number SFI/12/RC/2289_2, and by EU's Horizon 2020 research and innovation programs, under grant agreements n° 951911 AI4Media (https://ai4media.eu) and n° 825079 MindSpaces (https://mindspaces.eu/).

## References

1. Awad, G., Butt, A.A., Curtis, K., Fiscus, J.G., Godil, A., Lee, Y., Delgado, A., Zhang, J., Godard, E., Chocot, B., Diduch, L.L., Liu, J., Smeaton, A.F., Graham, Y., Jones, G.J.F., Kraaij, W., Quénot, G.: TRECVID 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In: Proc. TRECVID. Gaithersburg, MD, USA (2020)

2. Awad, G., Butt, A.A., Curtis, K., Lee, Y., Fiscus, J.G., Godil, A., Delgado, A., Zhang, J., Godard, E., Diduch, L.L., Smeaton, A.F., Graham, Y., Kraaij, W., Quénot, G.: TRECVID 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. In: Proc. TRECVID. Gaithersburg, MD, USA (2019)
3. Awad, G., Fiscus, J.G., Joy, D., Michel, M., Smeaton, A.F., Kraaij, W., Quénot, G., Eskevich, M., Aly, R., Ordelman, R., Ritter, M., Jones, G.J.F., Huet, B., Larson, M.A.: TRECVID 2016: Evaluating video search, video event detection, localization, and hyperlinking. In: Proc. TREC Video Retrieval Evaluation (TRECVID). Gaithersburg, MD, USA (2016)
4. Bates, M.J.: The design of browsing and berrypicking techniques for the online search interface. Online review **13**(5), 407–424 (1989)
5. Belkin, N.J., Marchetti, P.G., Cool, C.: BRAQUE: Design of an interface to support user interaction in information retrieval. Information Processing & Management **29**(3), 325–344 (1993)
6. Buckland, M.K.: On types of search and the allocation of library resources. Journal of the American Society for Information Science **30**(3), 143–147 (1979)
7. Christel, M.G.: Carnegie Mellon University traditional Informedia digital video retrieval system. In: Proc. ACM Conference on Image and Video Retrieval (CIVR). p. 647. Amsterdam, The Netherlands (2007)
8. Constantin, M.G., Hicks, S., Larson, M., Nguyen, N.T.: MediaEval multimedia evaluation benchmark: Tenth anniversary and counting. ACM SIGMM Records **12**(2) (2020)
9. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys **40**(2), 5:1–5:60 (2008)
10. Gurrin, C., Le, T.K., Ninh, V.T., Dang-Nguyen, D.T., Jónsson, B.Þ., Lokoč, J., Hürst, W., Tran, M.T., Schoeffmann, K.: Introduction to the Third Annual Lifelog Search Challenge (LSC'20). In: Proc. ACM International Conference on Multimedia Retrieval (ICMR). pp. 584–585 (2020)
11. Gurrin, C., Schoeffmann, K., Joho, H., Leibetseder, A., Zhou, L., Duane, A., Dang-Nguyen, D.T., Riegler, M., Piras, L., Tran, M.T., et al.: Comparing approaches to interactive lifelog search at the Lifelog Search Challenge (LSC2018). ITE Transactions on Media Technology and Applications **7**(2), 46–59 (2019)
12. Heller, S., Sauter, L., Schuldt, H., Rossetto, L.: Multi-stage queries and temporal scoring in vitrivr. In: Proc. IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 1–5 (2020)
13. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: A survey on visual content-based video indexing and retrieval. IEEE Transactions on Systems, Man, and Cybernetics, Part C **41**(6), 797–819 (2011)
14. Kraaij, W., Smeaton, A.F., Over, P., Arlandis, J.: TRECVID 2004 - an overview. In: Proc. TRECVID. Gaithersburg, MD, USA (2004)
15. Larson, M.A., Soleymani, M., Gravier, G., Ionescu, B., Jones, G.J.F.: The benchmarking initiative for multimedia evaluation: MediaEval 2016. IEEE MultiMedia **24**(1), 93–96 (2017)
16. Lee, J.H., Renear, A., Smith, L.C.: Known-item search: Variations on a concept. In: Proc. ASIS&T Annual Meeting). Austin, TX, USA (2006)
17. Li, X., Xu, C., Yang, G., Chen, Z., Dong, J.: W2VV++: Fully deep learning for ad-hoc video search. In: Proc. ACM Multimedia. Nice, France (2019)
18. Li, Y., Belkin, N.J.: A faceted approach to conceptualizing tasks in information seeking. Information processing & management **44**(6), 1822–1837 (2008)

19. Lokoč, J., Bailer, W., Schoeffmann, K., Münzer, B., Awad, G.: On influential trends in interactive video retrieval: Video Browser Showdown 2015–2017. IEEE Transactions on Multimedia **20**(12), 3361–3376 (2018)
20. Lokoč, J., Souček, T., Veselý, P., Mejzlík, F., Ji, J., Xu, C., Li, X.: A W2VV++ case study with automated and interactive text-to-video retrieval. In: Proc. ACM Multimedia. p. 2553–2561. Virtual Event / Seattle, WA, USA (2020)
21. Over, P., Awad, G., Jonathan Fiscus, B.A., Michel, M., Smeaton, A., Kraaij, W., Quénot, G.: TRECVID 2010 – An overview of the goals, tasks, data,evaluation mechanisms, and metrics. In: Proc. TRECVID. Gaithersburg, MD, USA (2010)
22. Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, G., Shaw, B., Smeaton, A.F., Kraaij, W., Quéot, G.: TRECVID 2013 - An overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proc. TRECVID. Gaithersburg, MD, USA (2013)
23. Rossetto, L., Bailer, W., Bernstein, A.: Considering human perception and memory in interactive multimedia retrieval evaluations. In: Proc. International Conference on Multimedia Modeling (MMM). pp. 605–616. Prague, Czech Republic (2021)
24. Rossetto, L., Gasser, R., Heller, S., Parian-Scherb, M., Sauter, L., Spiess, F., Schuldt, H., Peška, L., Souček, T., Kratochvíl, M., Mejzlík, F., Veselý, P., Lokoč, J.: On the user-centric comparative remote evaluation of interactive video search systems. IEEE MultiMedia (2021)
25. Rossetto, L., Gasser, R., Sauter, L., Bernstein, A., Schuldt, H.: A system for interactive multimedia retrieval evaluations. In: Proc. International Conference on MultiMedia Modeling (MMM). Prague, Czech Republic (2021)
26. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley (1989)
27. Schoeffmann, K.: A user-centric media retrieval competition: The Video Browser Showdown 2012-2014. IEEE MultiMedia **21**(4), 8–13 (2014)
28. Schoeffmann, K.: Video Browser Showdown 2012-2019: A review. In: Proc. International Conference on Content-Based Multimedia Indexing (CBMI). pp. 1–4. Dublin, Ireland (2019)
29. Smeaton, A.F., Kraaij, W., Over, P.: TRECVID 2003 - An overview. In: Proc. TRECVID. Gaithersburg, MD, USA (2003)
30. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVID. In: Proc ACM International Workshop on Multimedia Information Retrieval (MIR). pp. 321–330. Santa Barbara, California, USA (2006)
31. Snoek, C.G., Worring, M., de Rooij, O., van de Sande, K.E., Yan, R., Hauptmann, A.G.: VideOlympics: Real-time evaluation of multimedia retrieval systems. IEEE MultiMedia **15**(1), 86–91 (2008)
32. Walker, G., Janes, J.: Online retrieval: A dialogue of theory and practice. Libraries Unlimited (1999)
33. Wildemuth, B.M., O'Neill, A.L.: The "known" in known-item searches: Empirical support for user-centered design. College & Research Lib. **56**(3), 265–281 (1995)
34. Worring, M., Snoek, C., de Rooij, O., Nguyen, G., van Balen, R., Koelma, D.: Mediamill: Advanced browsing in news video archives. In: Proc. ACM Conference on Image and Video Retrieval (CIVR). pp. 533–536. Tempe, AZ, USA (2006)
35. Wu, J., Ngo, C.W.: Interpretable embedding for ad-hoc video search. In: Proc. ACM Multimedia. p. 3357–3366. Virtual Event / Seattle, WA, USA (2020)
36. Wu, J., Nguyen, P.A., Ma, Z., Ngo, C.W.: SQL-like interpretable interactive video search. In: Proc. International Conference on MultiMedia Modeling (MMM). pp. 391–397. Prague, Czech Republic (2021)